SANDIA REPORT

SAND2006-1706
Unlimited Release
Printed April 2006

Verification of LHS Distributions

Laura P. Swiler

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND06-1706
Unlimited Release
Printed March 2006

Verification of LHS Distributions

Laura P. Swiler
Optimization and Uncertainty Estimation Department
Sandia National Laboratories
PO Box 5800
Albuquerque, NM 87185-0370

Abstract

This document provides verification test results for normal, lognormal, and uniform distributions that are used in Sandia's Latin Hypercube Sampling (LHS) software. The purpose of this testing is to verify that the sample values being generated in LHS are distributed according to the desired distribution types. The testing of distribution correctness is done by examining summary statistics, graphical comparisons using quantile-quantile plots, and format statistical tests such as the Chi-square test, the Kolmogorov-Smirnov test, and the Anderson-Darling test. The overall results from the testing indicate that the generation of normal, lognormal, and uniform distributions in LHS is acceptable.

Acknowledgments

Ronald L. Iman, Michael J. Shortencarier, J. M. Davenport, and D. K. Ziegler developed the original LHS package at Sandia during the late 1970s and early 1980s (SAND83-2365). The authors are indebted to them for their pioneering work in the area of Latin hypercube sampling. Gregory Wyss and Sharon Daniel implemented a major upgrade to the software in the mid-1990s, converting it from Fortran 77 to Fortran 90, adding more than 25 new distributions, and including functionality that made the code much more portable (SAND98-0210). Michael Eldred, Sharon Daniel, Laura Swiler, and Shannon Brown have ported the 1998 version of LHS to a Linux/UNIX environment. The version of LHS that is in DAKOTA is documented in SAND2004-2439, "A User's Guide to Sandia's Latin Hypercube Sampling Software: LHS UNIX Library/Standalone Version."

Contents

Figures and Tables

1. Introduction

This document provides verification test results for the normal, lognormal, and uniform distributions that are used in Sandia's Latin Hypercube Sampling (LHS) software. The purpose of this testing is to verify that the sample values being generated in LHS are distributed according to the desired distribution types. The testing of distribution correctness is done by examining summary statistics, graphical comparisons using quantile-quantile plots, and format statistical tests such as the Chi-square test, the Kolmogorov-Smirnov test, and the Anderson-Darling test.

This document supports the Advanced Simulation and Computing (ASC) program's Verification and Validation (V&V) milestones. Many milestones use DAKOTA to perform uncertainty quantification (UQ) studies. The goal of uncertainty quantification is to understand the effect input uncertainties have on the uncertainty of the output, usually called a performance measure or measure of interest. A common method of performing UQ involves the following steps:

1. Assume certain distributions on the uncertain input variables or input parameters
2. Sample from those distributions
3. Run the simulation model (e.g., a finite element code) with the sampled values
4. Repeat Steps 1-3 with different sample draws to build up a distribution of the outputs.

In practice, one needs to have a method for generating random samples from specified distributions. At Sandia, we often use Latin Hypercube Sampling to generate samples. For many ASC codes, we use DAKOTA to perform UQ. This is done by calling the UNIX version of LHS from within DAKOTA. DAKOTA is a software toolkit which can call simulation models iteratively to perform various types of analysis such as uncertainty quantification, reliability analysis, parameter studies, and optimization studies. To perform an uncertainty quantification study in DAKOTA, one specifies the distributions on the input parameters of interest, tells DAKOTA to use LHS, and then DAKOTA does the overhead of calling the LHS code, getting the generated sample values from LHS, sending these values to the simulation model, and waiting for the simulation model to return the corresponding output values. DAKOTA allows for parallel execution of samples if desired, and DAKOTA collates the results and outputs various statistical measures of interest such as moments and percentiles of the output distribution, correlations between inputs and outputs, etc.

This verification study is not focused on DAKOTA. Rather it is focused on the specific version of LHS that is implemented within DAKOTA. The LHS code has a long pedigree and background, as explained in the next section. The version of LHS that is implemented in DAKOTA is what we refer to as the LHS UNIX Library/Standalone version because it may be called in library mode from DAKOTA or it may be called as a standalone code. The UNIX version is very similar to the latest LHS PC version developed by Greg Wyss, Sharon Daniel, and Kelly Jorgenson outlined in SAND98-0210 (Wyss and Jorgensen, 1998). However, a group of DAKOTA developers including Michael Eldred (the PI of the DAKOTA project), Laura Swiler, and Shannon Brown have made some modifications to the code to port it to the UNIX/Linux environment and make it more portable. LHS has undergone much testing over the years, and its widespread use has resulted in a

lot of distributed testing. Also, Wyss and Jorgenson did some testing of the LHS distributions in 1993. However, these test results were not formally written up and since the code has undergone some revisions with the porting to a UNIX environment, we thought it best to have a fresh start for the ASC milestones. DAKOTA has many advanced capabilities for UQ, including analytic reliability methods, stochastic finite element, and optimization under uncertainty. However, most users start with LHS for UQ: LHS is a core capability. LHS is used from within DAKOTA to support many of the ASC UQ milestones. This is the rationale for performing the verification studies on LHS.

The outline of this report is as follows: Section 2 provides information about the pedigree and background of the LHS code. Section 3 provides details about the verification and comparison methods used. Section 4 provides results for the normal distribution, Section 5 provides results for the lognormal distribution, and Section 6 provides results for the uniform distribution. Section 7 provides the results from some large scale testing, and Section 8 is the summary.

2. LHS Pedigree and Background

For more than twenty years, the Latin hypercube sampling (LHS) program has been successfully used to generate multivariate samples of statistical distributions. Its ability to use either Latin hypercube sampling with both random and restricted pairing methods has made it an important part of uncertainty analyses in areas ranging from probabilistic risk assessment (PRA) to complex simulation modeling.

Latin hypercube sampling was developed to address the need for uncertainty assessment for a particular class of problems. Consider a variable Y that is a function of other variables X_1, X_2, …, X_k. This function may be very complicated, for example, a computer model. A question to be investigated is "How does Y vary when the Xs vary according to some assumed joint probability distribution?" Related questions are "What is the expected value of Y?" and "What is the 99th percentile of Y?"

A conventional approach to these questions is to apply Monte Carlo sampling. By sampling repeatedly from the assumed joint probability density function of the Xs and evaluating Y for each sample, the distribution of Y, along with its mean and other characteristics, can be estimated. This approach yields reasonable estimates for the distribution of Y if the value of n is quite large. However, since a large value of n requires a large number of computations from the function or program of interest, which is potentially a very large computational expense, additional methods of uncertainty estimation were sought.

An alternative approach, which can yield more precise estimates, is to use a constrained Monte Carlo sampling scheme based on the idea of sample stratification. One such scheme, developed by McKay, Conover, and Beckman (1979), is Latin Hypercube Sampling. LHS selects n different values from each of k variables X_1, … X_k in the following manner. The range of each variable is divided into n nonoverlapping intervals on the basis of equal probability. One value from each

interval is selected at random with respect to the probability density in the interval. The n values thus obtained for X_1 are paired in a random manner (equally likely combinations) with the n values of X_2. These n pairs are combined in a random manner with the n values of X_3 to form n triplets, and so on, until n k-tuplets are formed. This is the Latin hypercube sample. It is convenient to think of this sample (or any random sample of size n) as forming an $(n \times k)$ matrix of input where the i^{th} row contains specific values of each of the k input variables to be used on the i^{th} run of the computer model. For more information about the sampling method, see Wyss and Jorgenson (SAND98-0210) or Swiler and Wyss (SAND2004-2439).

The original version of LHS developed at Sandia National Laboratories was documented in SAND83-2365 (Iman and Shortencarier). This code was substantially revised, extended, and upgraded in the mid-1990s. Gregory Wyss, Sharon Daniel, and Kelly Jorgensen designed and implemented much of this upgrade to the LHS software, converting it from Fortran 77 to Fortran 90, adding more than 25 new distributions, and including functionality that made the code much more portable. The revised version also included development of a Windows-based user interface to assist the user with input preparation as well as a graphical output system to support plotting of distributions generated by LHS. The documentation of the capabilities of the revised LHS code is presented in SAND98-0210 (Wyss and Jorgensen, 1998).

Michael Eldred, Sharon Daniel, Laura Swiler, and Shannon Brown ported the 1998 version of LHS (which was primarily designed for a Windows platform) to a Linux/UNIX environment in 2003-2004. This process involved writing some additional functionality to allow the LHS code to be called as a library from within the DAKOTA software environment ("input-by-call" vs. input by file), as well as some changes to modernize the code and make it more compatible with the needs of advanced simulators (e.g., converting single precision variables to double precision). The version of LHS that runs under a Linux or UNIX operating system can be compiled to run in two ways: called as a library or as a standalone LHS code run with file input (SAND2004-2439).

The Latin Hypercube Sampling code has a long pedigree, as evidenced by its history outlined above. While code longevity does not directly imply anything about code verification, it is likely that significant problems with the distributions would be noticed over many years of use. The purpose of this document is to supplement any previous testing of the distributions done formally or informally, and provide written documentation of the test results.

3. Verification Test Methods

There are many methods available to compare sample values with the true underlying distribution. This section draws heavily on the *Simulation Modeling and Analysis* textbook by Law and Kelton [1991] as well as the information provided by the National Institute of Standards in their Engineering Statistics Handbook [NIST *e-Handbook*].

There are three approaches we take for verification of the LHS distributions:

1. Summary Statistics
2. Graphical Comparisons
3. Formal Statistical Tests

These are explained in more detail below. Much of the actual testing was done using the Minitab and JMP software packages.

Summary Statistics

One of the first things to look at when analyzing a set of sample values are summary statistics about the sample, including the mean, standard deviation, skewness, and kurtosis. The skewness measures the symmetry of the distribution, and the kurtosis measures the weight of the tails in the distribution. Quantile summaries, which list various percentiles of the distribution, are also useful in determining whether the underlying distribution is symmetric or skewed, identifying any outliers, etc. Histograms and box plots (which are graphical representation of the quantiles, usually the quartiles) are also useful to understand the shape and spread of the data.

Finally, in the case where the samples are generated from a known distribution, one can perform statistical tests of various hypothesis, such as does the sample population mean equal the "true" mean with some confidence level, etc.? Comparing statistical measures such as mean and variance with the "true" distribution does not test the correctness of the entire distribution, but it provides useful information in the initial phase of verification.

Graphical Comparisons

Although these are not "formal" tests, one of the most common ways of testing for normality involves a graphical comparison. A probability plot graphically shows the comparison between the empirical distribution function (ECDF) calculated from the data, and the CDF of the "true" distribution function. Given a set of data, $X_1, X_2, \ldots, X_N$, the i^{th} order statistic is denoted $X_{(i)}$. It is the i^{th} smallest of the sample values $X_1, X_2, \ldots, X_N$. If one has a set of ordered sample points $X_{(1)}, X_{(2)}, \ldots X_{(N)}$, the ECDF is defined as F_N , where

$$F_N(X_{(i)}) = \frac{i}{N}$$

This is the proportion of the sample values that are less than or equal to $X_{(i)}$.

The probability plot is a graph of the sample probability $F_N(X_{(i)})$ vs. the fitted distribution probability $F(X_{(i)})$. If these values are close, then the P-P plot will be approximately linear with an intercept of zero and a slope of 1. If the probabilities are plotted against each other, it is called a P-P plot. If the quantiles are plotted against each other, it is called a Q-Q plot. Sometimes the probability is plotted against the quantile as shown in Figure 1. In this case, the sample values fall very close to the line, indicating the data likely follows a normal distribution.

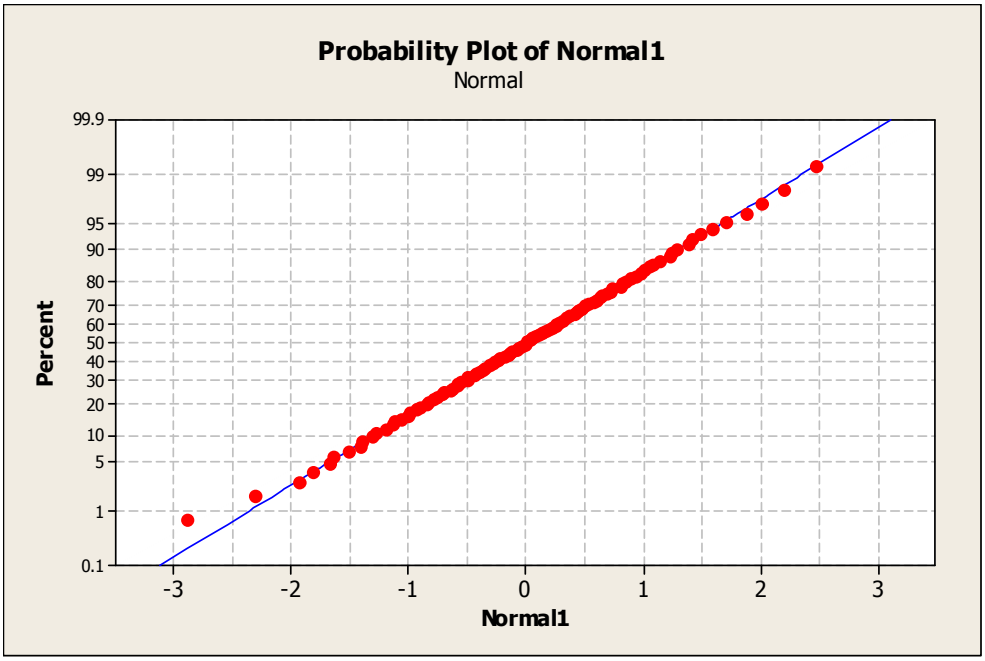


Figure 1. Example P-Q plot for the Normal Distribution

Formal Statistical Tests

There are a number of formal statistical tests, called "goodness-of-fit" tests. These tests are based on various types of hypotheses. The null hypothesis is:

> H_o: The sample data X_1, X_2, …,X_N are independently, identically distributed random variables with the distribution function F.

In practice, the hypotheses usually take the form: is a test statistic calculated from the sample data less than or greater than some threshold value, based on the distribution of the test statistic. If the test statistic calculated from the data is less than (or greater than, depending on the test), we can then accept the null hypothesis. However, Law and Kelton [1991] make an important point: failure to reject the null hypothesis should NOT be interpreted as accepting the null hypothesis is true. Some of these tests are not powerful for small sample sizes, and some tests are not very sensitive to small changes between the data and the fitted distribution. Thus, the tests are more useful for

detecting gross differences to a fitted distribution. Note also that most of these tests can be adapted for various distribution functions but some are specific to a particular distribution function. Finally, although most of the tests can be applied to a general distribution function, in practice, most of the statistical software packages have only implemented the tests for common distribution functions such as the normal distribution.

Chi-Square Tests

The oldest goodness-of-fit hypothesis test is the Chi-Square test. It involves a comparison between an empirical histogram based on the sample data and the density of the fitted distribution (the underlying distribution to which we are comparing). To calculate the Chi-Square test statistic, one divides the entire range of the fitted distribution into k intervals, for example $[a_0,a_1)$, $[a_1,a_2)$, ….$[a_{k-1},a_k)$. If N_j = the number of sample X values in the j^{th} interval $[a_{j-1},a_j)$, then the test statistic is:

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - Np_j)^2}{Np_j}$$

Where N is the total number of samples and p_j is the expected proportion of the sample values that should fall in the j^{th} interval: $p_j = \int_{a_{j-1}}^{a_j} f(X)dX$. Thus, the test statistic measures the normalized squared differences between the number we expect in each bin according to an underlying distribution, and how many sample values there actually are. The χ^2 test statistic should be small if the fit is good. The decision is to reject the null hypothesis if $\chi^2 > \chi^2_{k-1,1-\alpha}$, where k is the number of bins, and $\chi^2_{k-1,1-\alpha}$ is the upper 1-α critical value for a chi-square distribution with k-1 degrees of freedom. The critical value changes slightly if one estimates the parameters of the distribution from the data, which reduces the degrees of freedom.

The difficulty with implementing a χ^2 test is selecting the number and the size of the intervals. The test is sensitive to the choice of bins. Some approaches recommend choosing the intervals so that they are equiprobable: $p_1 = p_2 = \ldots p_k$. Additionally, to ensure validity of the test, there should be no intervals where the expected number in that interval is less than five. That is, $Np_j \geq 5 \forall j$. It is also recommended that the number of bins be at least three. This test works best with a large number of samples, and the test statistic is only valid at level α asymptotically as $N \rightarrow \infty$.

Kolmogorov-Smirnov

The Kolmogorov-Smirnov (K-S) test is used to detect if a sample population comes from a certain distribution. The K-S test is based on the empirical distribution function (ECDF) which was defined above. Recall that for a set of data, $X_1, X_2, \ldots, X_N$, the i^{th} order statistic is denoted $X_{(i)}$. It is the i^{th} smallest of the sample values $X_1, X_2, \ldots, X_N$.

The K-S test statistic is based on the difference between the empirical CDF and the "true" CDF. The K-S test statistic, D, is defined as :

$$D = \max_{(1 \le i \le N)} \left(F(X_{(i)}) - \frac{i-1}{N}, \frac{i}{N} - F(X_{(i)}) \right).$$

In this formula, $F(X_{(i)})$ is the theoretical cumulative distribution function (CDF) of the distribution against which we are trying to test. This distribution must be a continuous distribution, and its parameters must be specified and not estimated from the data. Note that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test for any number of points N, whereas the Chi-square test is valid in only asymptotically. The K-S test eliminates the need for binning data and specifying intervals as in the chi-square test. However, the K-S test does have limitations. It is mainly used for continuous distributions and is not easily applied to discrete distributions. It tends to be more sensitive near the center of the distribution than at the tails. And the distribution must be fully specified (e.g., the location, scale, and where appropriate, shape parameters of the distribution must be given). These parameters should NOT be estimated from the data. In recent years, the K-S test has been extended to allow for estimation of the parameters from the data. It was not possible to tell from the Minitab documentation how they are correcting the K-S test, since they are estimating the parameters.

Anderson-Darling test

One drawback to the K-S test is that it gives the same weight to the difference $|F_N(X) - F(X)|$ for every value of X. However, many distributions differ primarily in the tails. The Anderson-Darling test is designed to detect discrepancy in the tails and has higher power than the K-S test for many distributions.

The A-D test statistic A_n^2 is defined as:

$$A_n^2 = N \int_{-\infty}^{\infty} [F_N(X) - F(X)]^2 \psi(X) f(X) dX$$

Where $\psi(X)$ is the weight function $\psi(X) = \frac{1}{F(X)(1-F(X))}$. This means that A_n^2 is a weighted average of the squared differences $|F_N(X) - F(X)|^2$ and the weights are largest for the tails of F(X), where F(X) is close to zero or one. The form of the test is to reject the null hypothesis if A_n^2 exceeds some critical value that is a function of N and α. Tables of these critical values have been compiled for a few distributions, including the normal distribution.

Shapiro-Wilk/Ryan-Joiner

The Shapiro-Wilk and Ryan-Joiner test are very similar, and based on the correlation one would expect between the sample data set and the data one would expect if the underlying distribution

were normal. The test statistic, W, is constructed so that small values of W are evidence of departure from normality. The test statistic is:

$$W = \frac{(\sum_{i=1}^{N} a_i X_{(i)})^2}{\sum_{i=1}^{N} (X_i - X)^2}$$

where a_i are constants generated from the moments of the order statistics.

Statistical Software

Minitab offers three tests for normality: the Anderson-Darling test, the Ryan-Joiner test, and the Kolmogorov-Smirnov test. Minitab does not offer a test for uniform distributions, however it is possible to construct a Chi-square test statistic from the data and test that. JMP uses a Kolmogorov-Smirnov test to detect normality when the mean and variance of the fitted distribution are known; it uses a Shapiro-Wilks test when the mean and variance of the underlying distribution are not known. JMP also does not specifically have a test constructed for the uniform distribution. Both JMP and Minitab offer a variety of probability and quantile plots, as well as summary statistics about the sample data.

4. The Normal Distribution

The LHS software implemented in DAKOTA provides the user with two different methods for sampling from the normal distribution. The normal distribution is defined by the density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} \quad -\infty < x < \infty,$$

where the distribution mean and variance are μ and σ^2, respectively. The standard deviation of the distribution, which is required by LHS as an input parameter for several normal distribution sampling methods, is denoted by σ. The first sampling method for the normal distribution samples over all quantiles. The bounded normal method samples a normal distribution that is bounded.

For the purposes of the V&V analysis of the regular normal distribution in LHS, three runs of the LHS code were performed in DAKOTA. Each run involved 2 normally distributed, uncorrelated random variables. Each random variable was chosen from the standard normal distribution, with zero mean and standard deviation of one. The first run had 100 samples, the second run had 1000 samples, and the third run had 10,000 samples. The sections below provide results of testing with these sample data sets.

Summary Statistics: N = 100

For the 100 sample data sets, here are the results from Minitab:

Descriptive Statistics: 1n100

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
TF1n	100	-0.005	0.101	1.008	-2.887	-0.680	0.005	0.676	2.473

Descriptive Statistics: 2n100

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
TF2n	100	0.006	0.102	1.017	-2.655	-0.691	-0.000	0.670	3.010

For a N(0,1) distribution, we expect the sample mean and standard deviation to be approximately zero and one. For the two samples we took with N=100, we see this. We also expect the 25th and 75th percentiles to be -0.6745 and +0.6745 respectively. These percentiles are approximately correct. Finally, note that the maximum and minimum vary quite a bit.

To test if the mean of the sample data truly is zero, based on the assumption that the underlying distribution is normal, we can use a t-test, where:
H_o: $\mu = 0$; H_A: $\mu \neq 0$.

The hypothesis test is to accept H_o at significance level α if $|t^*| \le t(1-\alpha/2, n-1)$. For data set 1n100, we have: $|t^*| = \left|\frac{-.005}{0.101}\right| = 0.0495 \le t(1-\alpha/2, n-1) = 1.9842$. Thus, we accept the null hypothesis that the mean is equal to zero. The same conclusion can be made for data set 2n100.

To test if the variance of the sample data is one, based on the assumption that the underlying distribution is normal, we can use a Chi-square test, where:
$H_o : \sigma^2 = 1, H_A : \sigma^2 \neq 1$
The hypothesis test is to accept H_o at significance level α if $\chi^2(\alpha/2, n-1) \le \frac{(n-1)s^2}{\sigma_o^2} \le \chi^2(1-\alpha/2, n-1)$ where s^2 is the sample variance. For data set 1n100, we have: $\chi^2(\alpha/2, n-1) = 73.336 \le 100.59 \le 128.42 = \chi^2(1-\alpha/2, n-1)$ and thus we can accept the null hypothesis that the variance is equal to one. The same conclusion can be made for 2n100.

Graphical Comparisons: N = 100

The graphical comparisons with a quantile plot of the sample data (red points) vs. a normal distribution (blue line) in Figures 2 and 3 show the agreement is very good for both sample sets:

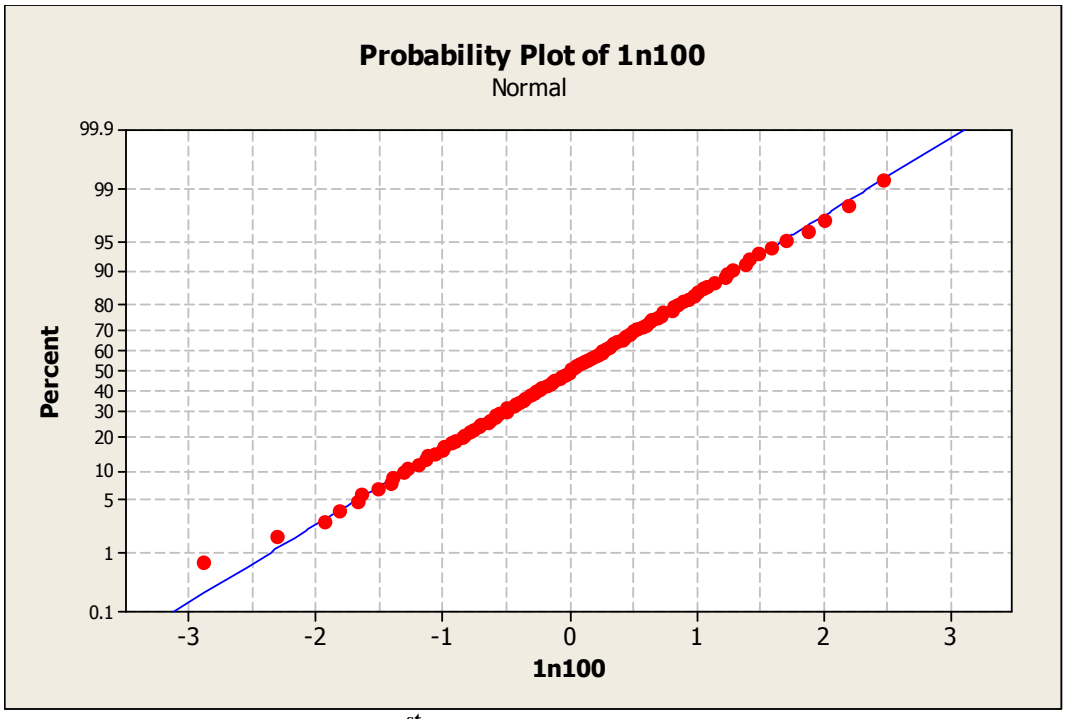


Figure 2. P-Q plot of the 1st Normal LHS sample, with sample size = 100

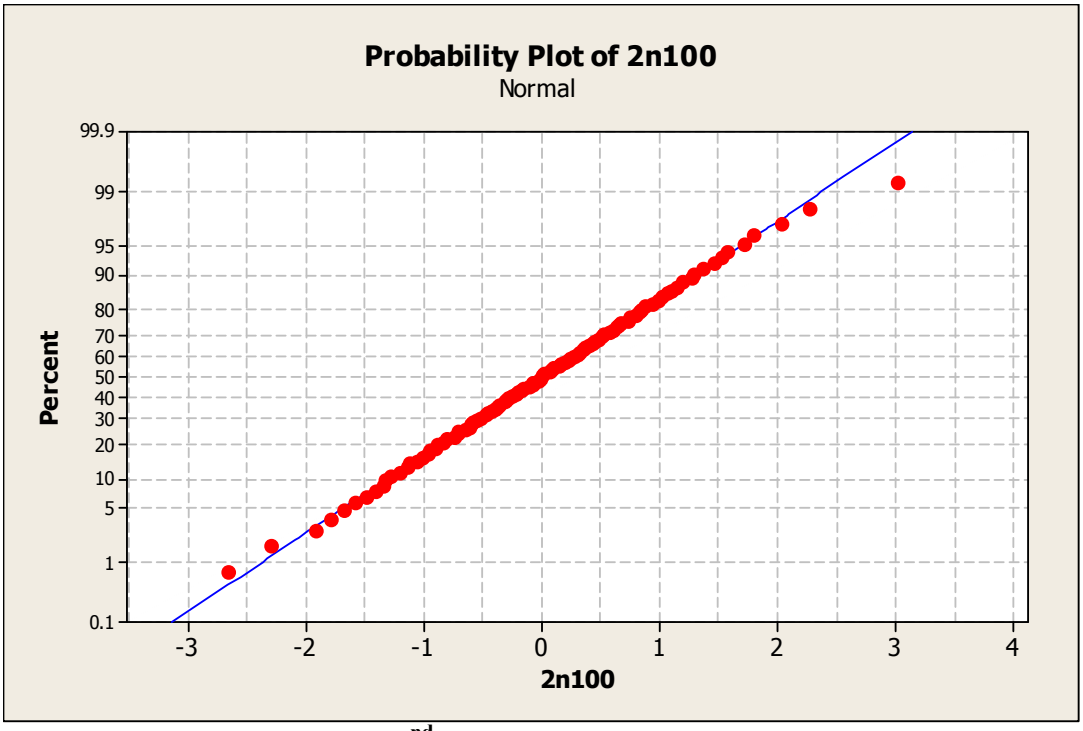


Figure 3. P-Q plot of the 2nd Normal LHS sample, with sample size = 100

Formal Tests: N = 100

The Anderson-Darling test statistic as calculated in Minitab for 1n100 is 0.025, with a p-value of 1.00. The interpretation of this is that if the p-value is less than the desired significance level α, then one must reject the null hypothesis. Otherwise, the null hypothesis is accepted. In this case, for $\alpha = 0.05$, we accept the null hypothesis that the data do follow a normal distribution. We also accept the Anderson-Darling test for the second data set of 100 points, 2n100, with a test statistic of 0.034.

The Kolmogorov-Smirnov test statistic as calculated in Minitab for 1n100 is 0.013. Minitab only gives a p-value for this test, and specifically the output is $p > 0.15$, meaning that p-value is greater than 0.15. Since we usually have an alpha value of 0.05 or 0.10, then we would accept the null hypothesis that the data do follow a normal distribution according to this test. The KS test statistic for 2n100 is 0.015. The p-value for the second data set was also $p>0.15$.

The Ryan-Joiner test resulted in accepting the null hypothesis for both data sets, but at a weaker level than the above two test. The Ryan-Joiner test statistic for both data sets was the same, a value of 0.999. The p-value in both cases was $p>0.10$. Thus, for an alpha value of 0.05, we would still accept the null hypothesis.

The results from JMP are shown in Figure 4. JMP produces much of the same output as Minitab does, in a different format. A histogram of each of the 2 samples is shown in green, with a probability density function for the "true" normal(0,1) overlaid in red. The quantile-quantile plot is shown at the top. The quantiles, moments, and confidence intervals for the mean and standard deviation are listed. Finally, a Kolmogorov-Smirnov goodness-of-fit test is performed. Note that the K-S test statistic is slightly different than that calculated in Minitab. In Minitab, the K-S test statistic is 0.013 and 0.015 for samples 1 and 2, respectively, while in JMP it is .00999 and 0.00998. This difference is due to the fact that Minitab is using the sample mean and standard deviation, while JMP is using the specified (0,1) mean and standard deviation. The difference may also be due to slight differences in the way people calculate the empirical distribution function: some approaches normalize it. Note that the last section in JMP states that the probability that the test statistic is greater than D is 25% in both cases. This means that the probability of obtaining a greater test-statistic value D by chance alone is 25%. To see if D is significant, we can use the approximation given in [Law and Kelton]: reject H_o if: $\left(\sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}}\right) D > c_{1-\alpha}$, where the value of $c_{1-\alpha}$ is 1.38 when $\alpha = 0.05$. In the case of sample 1, this test reduces to: $10.131 * 0.009998 = 0.1013 > 1.38$ which is not true, so we do not reject the null hypothesis that this data comes from a normal distribution.

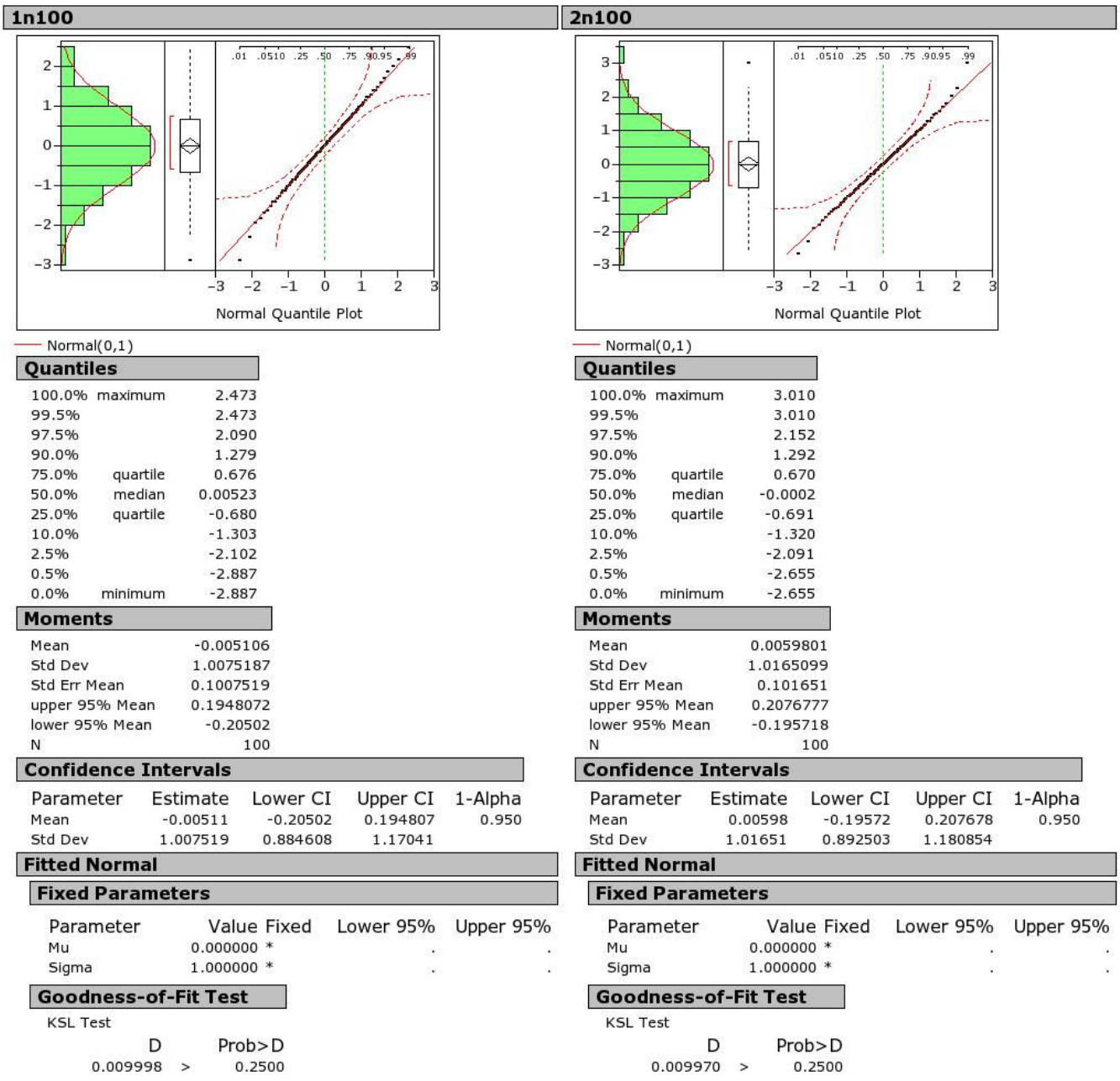


Figure 4. Summary Statistics, Normal Distribution, N = 100

Summary Statistics: N = 1000

For the 1000 sample data sets, here are the results from Minitab:

Descriptive Statistics: 1n1000, 2n1000

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
1n1000	1000	-0.00047	0.0316	1.0005	-3.5476	-0.6748	-0.0001	0.6760	3.0986
2n1000	1000	-0.00004	0.0316	1.0003	-3.2397	-0.6751	0.0001	0.6747	3.2419

We see that the mean and standard deviation are closer to (0,1) than the values obtained from the 100 point sample sets. Also, the 25th and 75th percentiles are very close to the expected values of -0.6745 and +0.6745 respectively.

As before, we use a t-test to test if the mean is zero: H_o: $\mu = 0$; H_A: $\mu \neq 0$.

The hypothesis test is to accept H_o at significance level α if $|t^*| \le t(1-\alpha/2, n-1)$. For data set 1n1000, we have: $|t^*| = \left|\frac{-.00047}{0.0316}\right| = 0.0149 \le t(1-\alpha/2, n-1) = 1.9842$. Thus, we accept the null hypothesis that the mean is equal to zero. The same conclusion can be made for data set 2n1000.

To test if the variance of the sample data is one, based on the assumption that the underlying distribution is normal, we can use a Chi-square test, where: $H_o : \sigma^2 = 1, H_A : \sigma^2 \neq 1$.

The hypothesis test is to accept H_o at significance level α if $\chi^2(\alpha/2, n-1) \le \frac{(n-1)s^2}{\sigma_o^2} \le \chi^2(1-\alpha/2, n-1)$ where s^2 is the sample variance. For data set 1n1000, we have: $\chi^2(\alpha/2, n-1) = 73.336 \le 99.099 \le 128.42 = \chi^2(1-\alpha/2, n-1)$ and thus we can accept the null hypothesis that the variance is equal to one. The same conclusion can be made for 2n1000.

Graphical Comparisons: N = 1000

The graphical comparisons with a quantile plot of the sample data (red points) vs. a normal distribution (blue line) in Figures 5 and 6 show the agreement is very good for both sample sets:

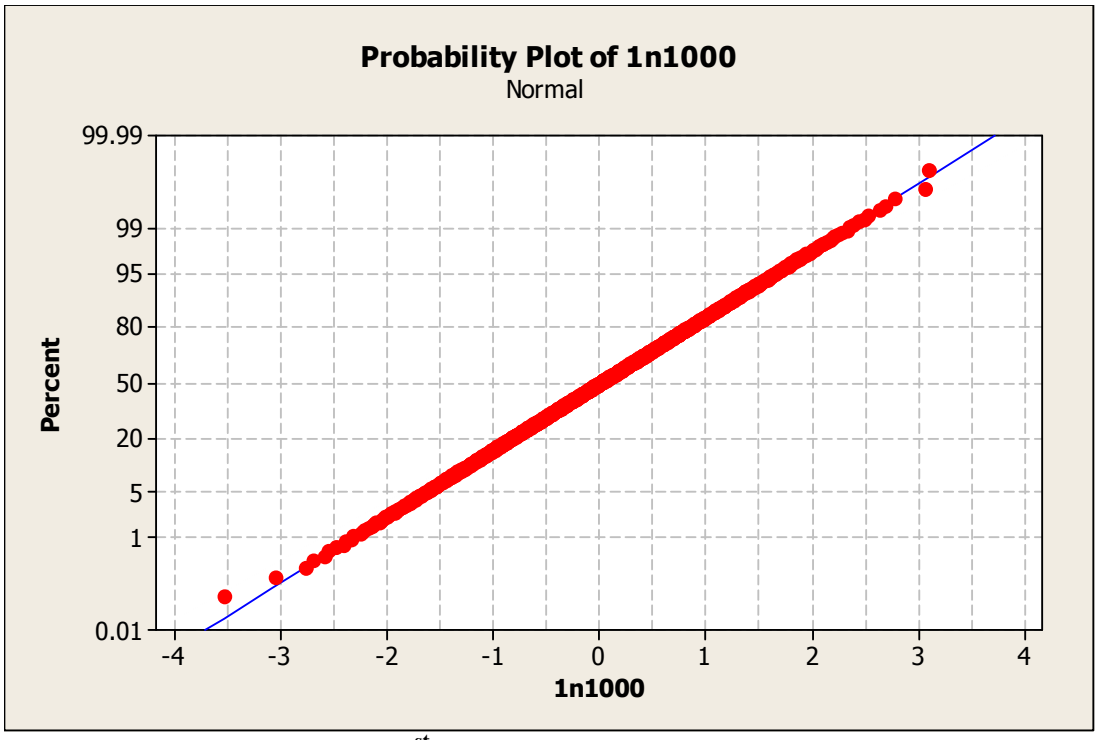


Figure 5. P-Q plot of the 1st Normal LHS sample, with sample size = 1000

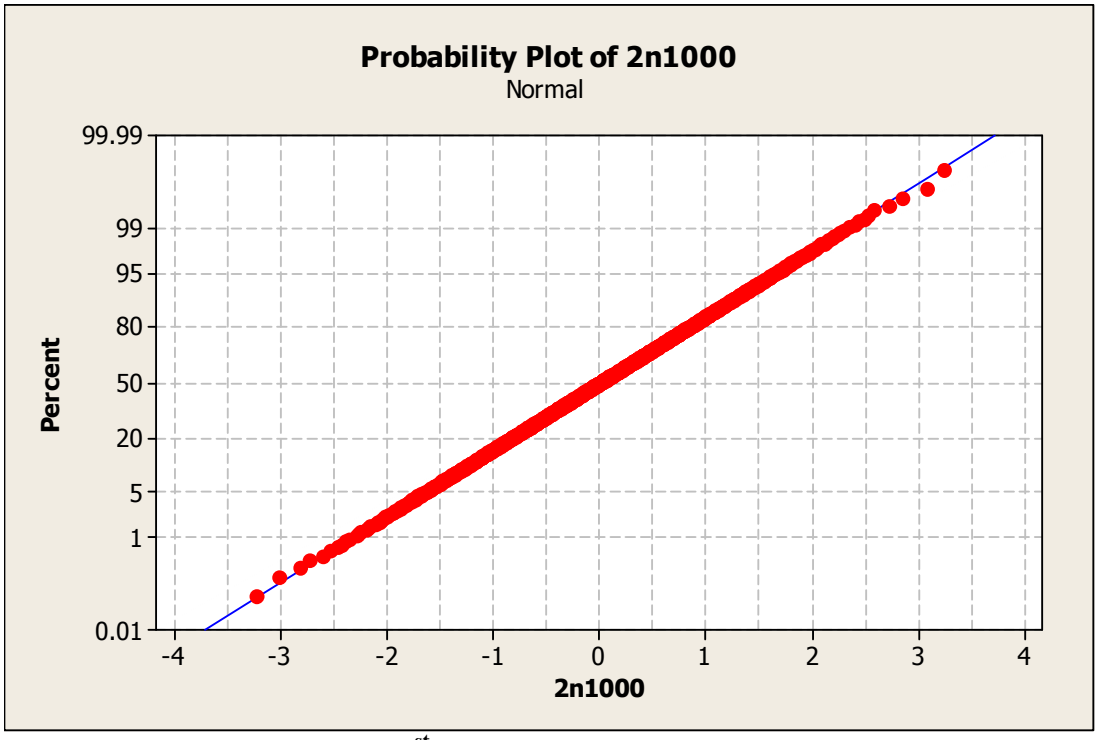


Figure 6. P-Q plot of the 1st Normal LHS sample, with sample size = 1000

Formal Tests: N = 1000

The Anderson-Darling test statistic as calculated in Minitab for 1n1000 is 0.004, with a p-value of 1.00. The interpretation of this is that if the p-value is less than the desired significance level α, then one must reject the null hypothesis. Otherwise, the null hypothesis is accepted. In this case, for $\alpha = 0.05$, we accept the null hypothesis that the data do follow a normal distribution. We also accept the Anderson-Darling test for the second data set of 1000 points, 2n1000, with a A-D test statistic of 0.003 with a p-value of 1.00.

The Kolmogorov-Smirnov test statistic as calculated in Minitab for 1n1000 is 0.001. Minitab only gives a p-value for this test, and specifically the output is $p > 0.15$, meaning that p-value is greater than 0.15. Since $\alpha = 0.05$, then we would accept the null hypothesis that the data do follow a normal distribution according to this test. The KS test statistic for 2n1000 is also 0.001, with $p>0.15$.

The Ryan-Joiner test resulted in accepting the null hypothesis for both data sets. The Ryan-Joiner test statistic for both data sets was the same, a value of 1.0. The p-value in both cases was $p>0.10$. Thus, for an alpha value of 0.05, we would still accept the null hypothesis.

The results from JMP are shown in Figure 7. JMP produces much of the same output as Minitab does, in a different format. A histogram of each of the 2 samples is shown in green, with a probability density function for the "true" normal(0,1) overlaid in red. The quantile-quantile plot is shown at the top. The quantiles, moments, and confidence intervals for the mean and standard deviation are listed. Finally, a Kolmogorov-Smirnov goodness-of-fit test is performed. In this case, the K-S statistic is the same as that calculated in Minitab.
To see if D is significant, we can use the approximation given in [Law and Kelton]: reject H_o if: $\left(\sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}}\right) D > c_{1-\alpha}$, where the value of $c_{1-\alpha}$ is 1.38 when $\alpha = 0.05$. For both sample sets 1 and 2, we do not reject the null hypothesis that this data comes from a normal distribution.

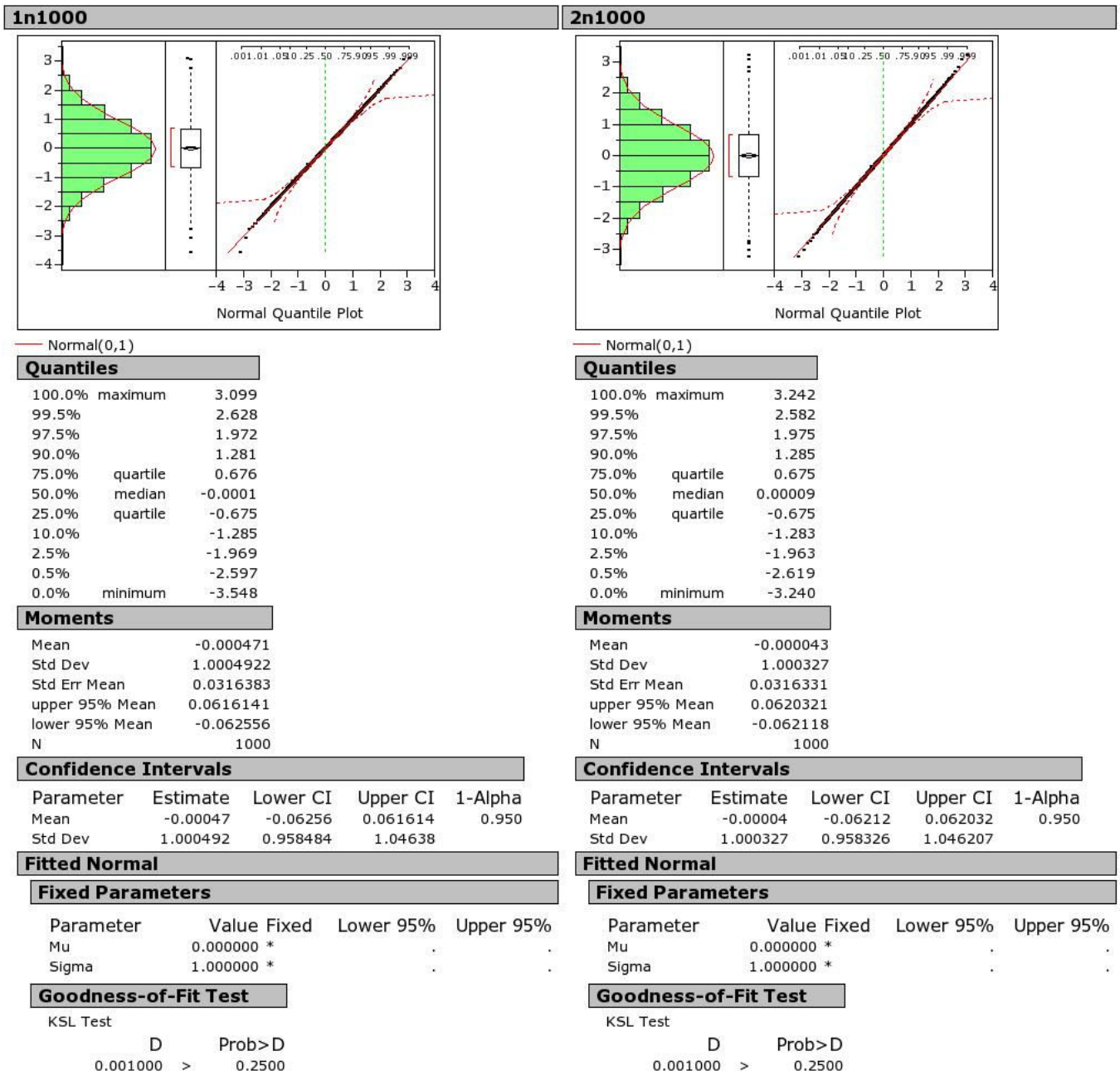


Figure 7. Summary Statistics, Normal Distribution, N = 1000

Summary Statistics: N = 10000

For the 10000 sample data sets, here are the results from Minitab:

Descriptive Statistics: 1n10000, 2n10000

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
1n10000	10000	-0.0000421	0.0100	1.0001	-4.1141	-0.6745	0.0001	0.6745	3.8440
2n10000	10000	-0.0000125	0.0100	1.0000	-3.8754	-0.6746	-0.0000	0.6745	3.7556

We see that the mean and standard deviations are closer to (0,1) than the values obtained from the 1000 point sample sets, as expected. Also, the 25th and 75th percentiles are very close to the expected values of -0.6745 and +0.6745 respectively.

As before, we use a t-test to test if the mean is zero: H_o: $\mu = 0$; H_A: $\mu \neq 0$.

The hypothesis test is to accept H_o at significance level α if $|t^*| \le t(1-\alpha/2, n-1)$. For data set 1n10000, we have: $|t^*| = \left|\frac{-.0000421}{0.01}\right| = 0.00421 \le t(1-\alpha/2, n-1) = 1.9842$. Thus, we accept the null hypothesis that the mean is equal to zero. The same conclusion can be made for data set 2n10000.

To test if the variance of the sample data is one, based on the assumption that the underlying distribution is normal, we can use a Chi-square test, where: $H_o : \sigma^2 = 1, H_A : \sigma^2 \neq 1$.

The hypothesis test is to accept H_o at significance level α if $\chi^2(\alpha/2, n-1) \le \frac{(n-1)s^2}{\sigma_o^2} \le \chi^2(1-\alpha/2, n-1)$ where s^2 is the sample variance. For data set 1n10000, we have: $\chi^2(\alpha/2, n-1) = 73.336 \le 99.02 \le 128.42 = \chi^2(1-\alpha/2, n-1)$ and thus we can accept the null hypothesis that the variance is equal to one. The same conclusion can be made for 2n10000.

Graphical Comparisons: N = 10000

The graphical comparisons with a quantile plot of the sample data (red points) vs. a normal distribution (blue line) in Figures 8 and 9 show the agreement is very good for both sample sets:

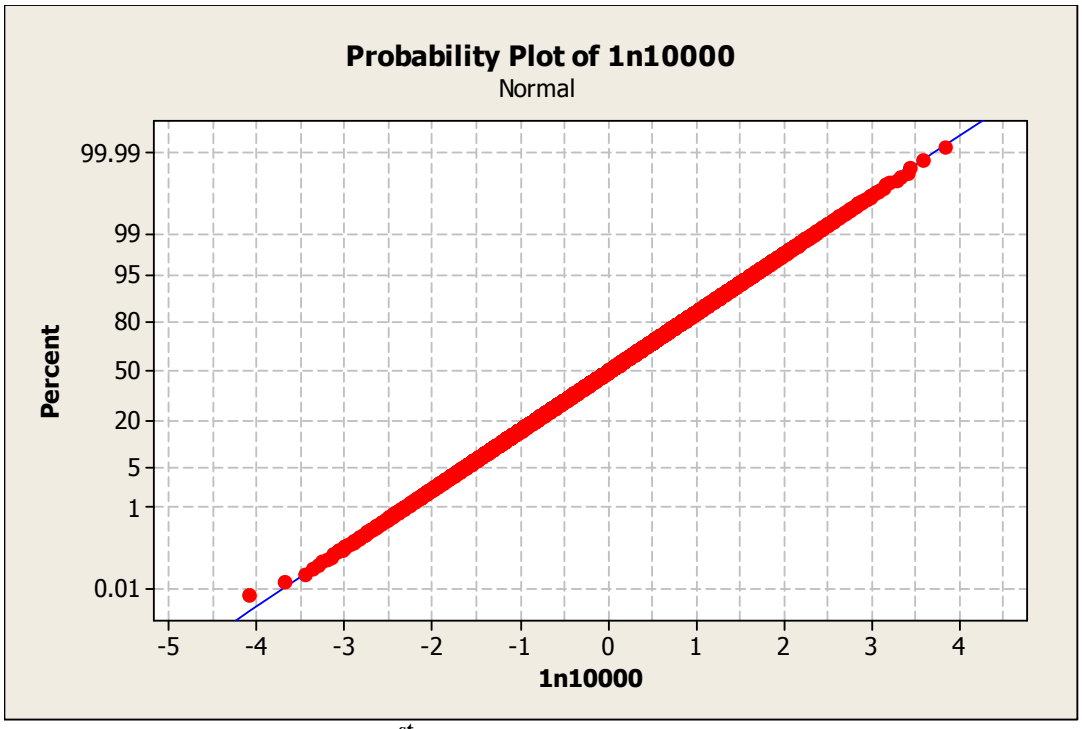


Figure 8. P-Q plot of the 1st Normal LHS sample, with sample size = 10000

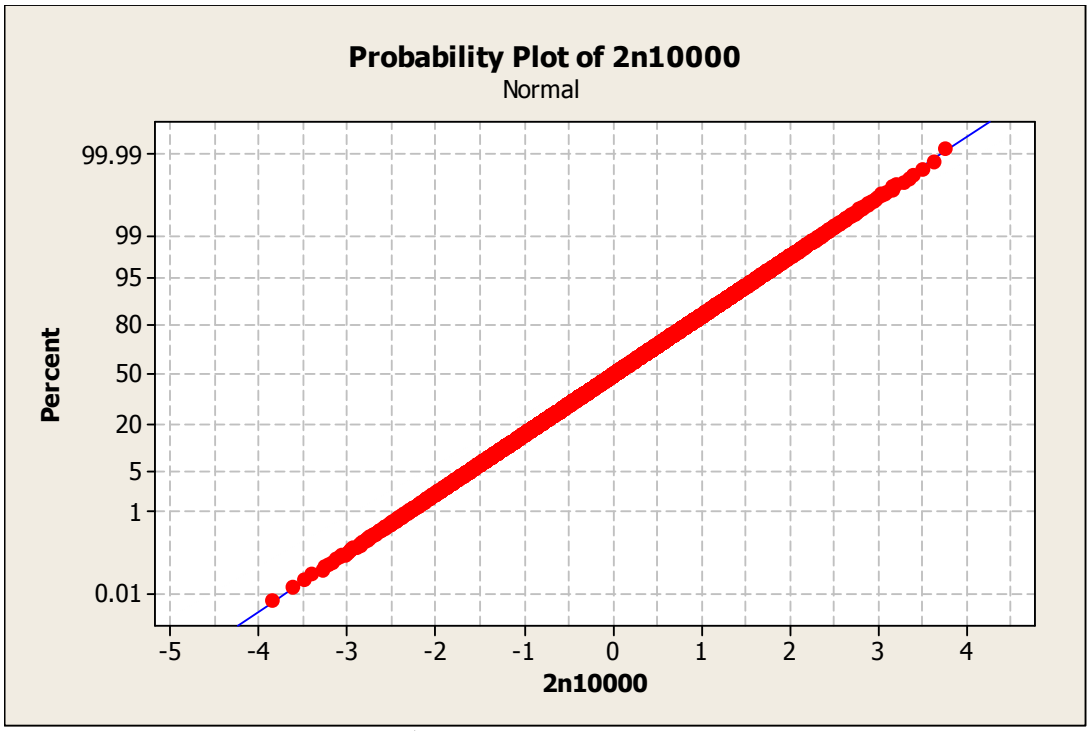


Figure 9. P-Q plot of the 1st Normal LHS sample, with sample size = 10000

Formal Tests: N = 10000

The Anderson-Darling test statistic as calculated in Minitab for 1n10000 is 0.000, with a p-value of 1.00. The interpretation of this is that if the p-value is less than the desired significance level α, then one must reject the null hypothesis. Otherwise, the null hypothesis is accepted. In this case, for $\alpha = 0.05$, we accept the null hypothesis that the data do follow a normal distribution. We also accept the Anderson-Darling test for the second data set of 10000 points, 2n10000, with a A-D test statistic of 0.000 with a p-value of 1.00.

The Kolmogorov-Smirnov test statistic as calculated in Minitab for 1n10000 and 2n10000 is zero. Minitab only gives a p-value for this test, and specifically the output for both sample sets is $p > 0.15$, meaning that p-value is greater than 0.15. Since $\alpha = 0.05$, then we would accept the null hypothesis that the data do follow a normal distribution according to this test.

The Ryan-Joiner test resulted in accepting the null hypothesis for both data sets. The Ryan-Joiner test statistic for both data sets was the same, a value of 1.0. The p-value in both cases was $p > 0.10$. Thus, for an alpha value of 0.05, we would still accept the null hypothesis.

The results from JMP are shown below in Figure 10. Note that with 10000 samples, the histogram (in green) is extremely close to the true distribution, when compared with the true normal density (in red). Also note that the confidence limits about the mean and standard deviation are very tight, as to be expected with such a large number of samples. However, the true mean and standard deviation lie within the 95% confidence intervals. Finally, the KS test statistic is very small, 0.0001 for both samples. Again, using the approximation to see if D is significant, we would reject H_o if: $\left(\sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}}\right) D > c_{1-\alpha}$, where the value of $c_{1-\alpha}$ is 1.38 when $\alpha = 0.05$. For both sample sets 1 and 2, we do not reject the null hypothesis that this data comes from a normal distribution.

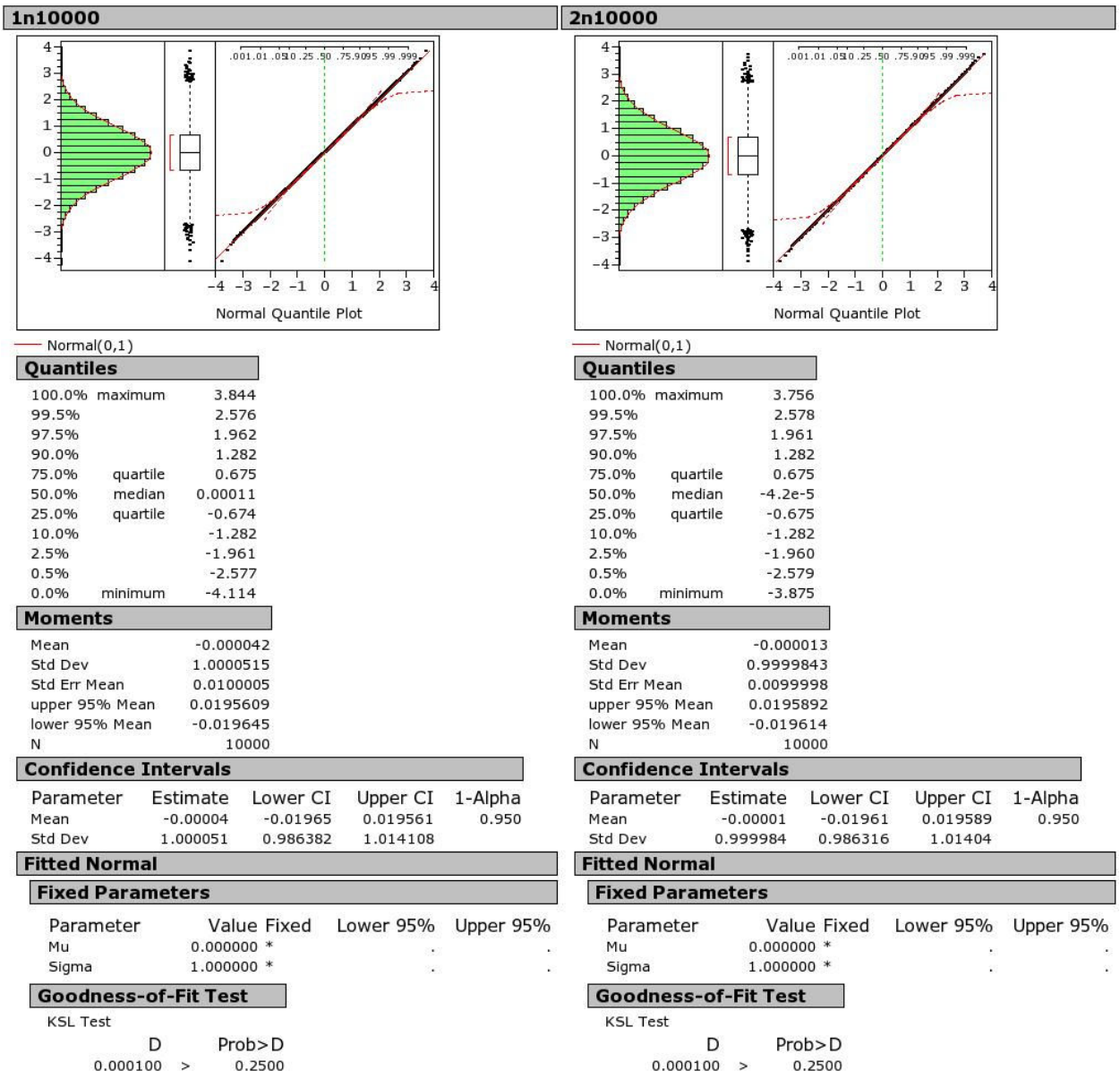


Figure 10. Summary Statistics, Normal Distribution, N = 10000

Table 1 summarizes the results of testing the normal distribution, showing that in all cases we cannot reject the hypothesis of the underlying distribution being a N(0,1) distribution:

Test Statistic for Normal Distribution	**Test Statistic Value**		**Reject null hypothesis?**
N = 100	Sample 1	Sample 2	
Anderson-Darling	0.025	0.034	NO
Kolmogorov-Smirnov	0.013	0.015	NO
Ryan-Joiner	0.999	0.999	NO
N = 1000			
Anderson-Darling	0.004	0.003	NO
Kolmogorov-Smirnov	0.001	0.001	NO
Ryan-Joiner	1.0	1.0	NO
N=10000			
Anderson-Darling	0	0	NO
Kolmogorov-Smirnov	0.0001	0.0001	NO
Ryan-Joiner	1.0	1.0	NO

Table 1. Summary Statistics, Normal Distribution

5. The Lognormal Distribution

The LHS software implemented in DAKOTA provides the user with two different methods for sampling from the lognormal distribution. The lognormal distribution is a distribution whose logarithm is described by a normal distribution. The lognormal distribution is defined by the density function:

$$f(x) = \frac{1}{x\sigma_N\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu_N)^2}{2\sigma^2} \right] \quad -\infty < x < \infty,$$

where the mean and variance of the underlying normal distribution are μ_N and σ_N^2, respectively. In DAKOTA, the user is required to enter the mean and either the standard deviation or error factor for the lognormal distribution. These are related to the underlying normal distribution parameters by the following formulas:

$$\mu_{LN} = e^{(\mu_N + \sigma_N^2)}, \; \sigma_{LN}^2 = e^{2\mu_N + \sigma_N^2}(e^{\sigma_N^2} - 1.0), \; \varepsilon_{LN} = e^{(1.645*\sigma_N)}$$

Where μ_{LN} is the mean of the lognormal distribution, σ_{LN}^2 is the variance, and ε_{LN} is the "error factor" which is defined as the ratio of the 95th percentile to the median of the lognormal distribution.

For the purposes of the V&V analysis of the lognormal distribution in LHS, three runs of the LHS code were performed in DAKOTA. Each run involved 2 lognormally distributed, uncorrelated random variables. Each random variable was chosen so that the underlying normal distribution was the standard normal distribution, with zero mean and standard deviation of one. This translates to a lognormal mean of 1.647821, and a lognormal standard deviation of 2.161197. The DAKOTA lognormal input specification using the error factor instead of the standard deviation was also tested and produced nearly identical results to the samples generated with the standard deviation specified. We present only the results using the standard deviation specification. As with the normal distribution, the first run had 100 samples, the second run had 1000 samples, and the third run had 10,000 samples. The sections below provide results of testing with these sample data sets.

Note that the results here are presented a little differently than the results of the normal distribution presented in Section 3. For each set of samples, we present both the raw data (the lognormal distribution) and then we take the log of the samples and present the underlying normal distribution based on the sample data. The formal statistical tests that are specific to the normal distribution can then be applied to the log-transformed data.

Summary Statistics and Graphical Comparison: N = 100

Figure 11 shows the raw data based on the samples generated by LHS, and Figure 12 shows the log-transformed data. There is one very important point to remember when examining this data. The lognormal distribution has a long tail. In the data we generated with a lognormal mean of 1.648 and a lognormal standard deviation of 2.161, the 99th percentile value of this distribution is 10.42. This means that in 100 samples generated by LHS, only one will lie in the bin from [10.42, ∞] because of the way the stratification is done. Thus, in these first two sample sets, we see a large difference in the maximum values: Sample set 1ln100 had a maximum value of 11.856, while sample set 2ln100 had a maximum value of 20.297. The difference in the maximum values greatly affects the variance and standard deviation of the sample sets. The true value for the median of this distribution is 1.0. Both samples have medians very close to this. The true 75th percentile value is 1.963. Both samples are close to this, and likewise with the 90th percentile which has a true value of 3.602. However, statistics of these 100 sample sets are not as good at matching the true 97.5th percentile, which is 7.099, and the sample means and standard deviations are not very close to what was specified in the input specification (a lognormal mean of 1.647821, and a lognormal standard deviation of 2.161197).

The inability of the lognormal samples to match the specified means and standard deviations with 100 samples should not be of concern. With only 1% of the distribution lying in [10.42, ∞], there will be only one LHS sample taken in this interval and the location of that particular sample will strongly affect the mean and standard deviation. This does not mean that the sample generated is not lognormal: when we transform to normal space and do the formal statistical tests, we see that we cannot reject the hypothesis that the underlying distribution is normal. Furthermore, as we take more samples, we see these statistics converge to their true estimates. This result is due to the fact that we are sampling a very long-tailed distribution sparsely, and highlights the limitations of small sample numbers if one wants to estimate tail probabilities accurately.

Finally, note a few things: The Kolmogorov-Smirnov-Lillifors test shows that we cannot reject the null hypothesis of the data being lognormal. Also, note that the 95% confidence intervals around the means and standard deviations DO capture the true values for sample 1ln100. The true mean is captured in the second sample, 2ln100, but the standard deviation is not. However, these confidence intervals are based on the assumption that the underlying distribution is normal. Thus, they should not strictly be used in the case where we have a clearly non-normal distribution, but we can examine these confidence intervals as a sanity check.

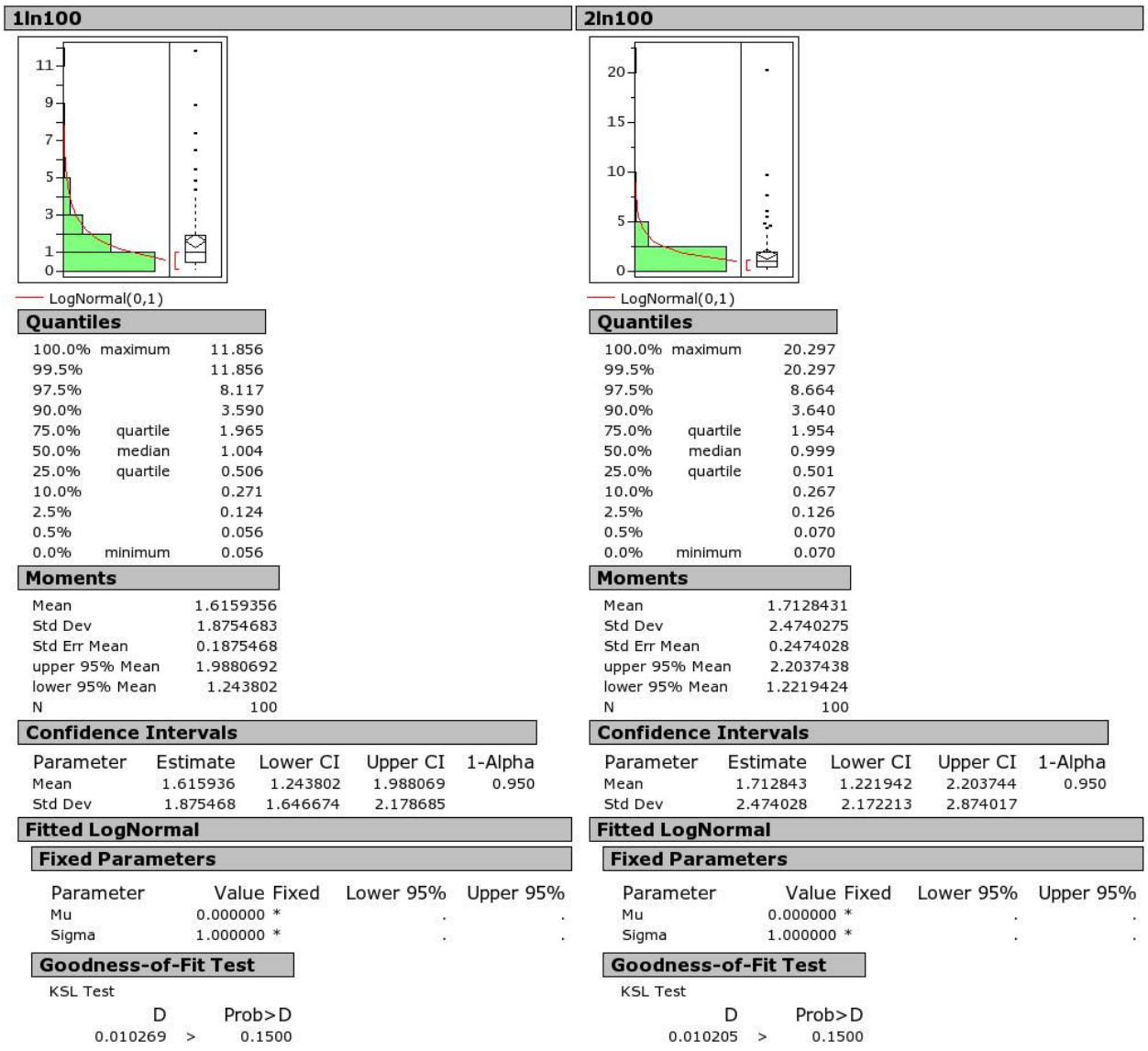


Figure 11. Summary Statistics, Lognormal Distribution, N = 100

Figure 12 shows the log-transformed data for the 100 sample sets:

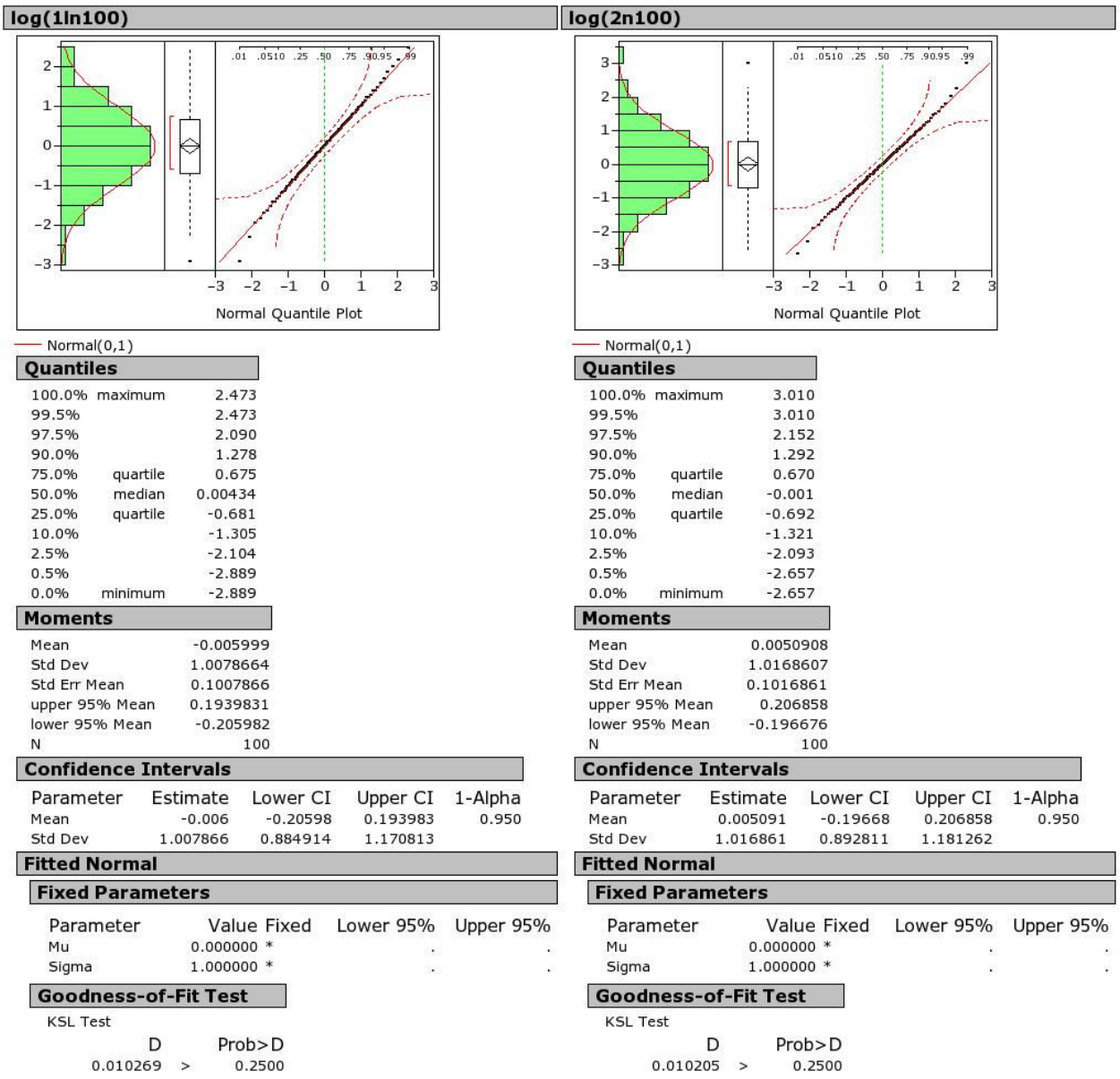


Figure 12. Summary Statistics, Log-transformed Lognormal Distribution, N = 100

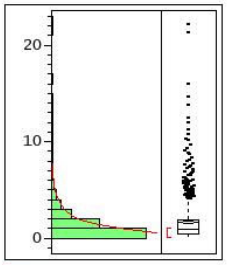
Note that the log transform of the lognormal data fits a normal distribution very well. In this case, we can use the confidence intervals because they are valid, and the true values for the mean and standard deviation (0,1) do fall within the confidence intervals in both sample sets. The quantile plots and the KSL test verify that we cannot reject the hypothesis that the underlying samples come from a normal distribution.

Summary Statistics and Graphical Comparison: N = 1000

We perform a similar transformation for N=1000: first we look at the raw data, then log-transform it. The raw data has a lognormal distribution shape, as shown in Figure 13. Note that in these samples, the maximum value is in the mid-twenties. The 99.5th percentile of the true distribution is 13.142. For 1000 samples, we expect 5 samples to lie above 13.142. Sample set 1ln1000 and sample set 2ln2000 both have exactly 5 samples above this value. The 97.5th percentile estimate is more accurate than that generate in the 100-point sample sets, as expected. Most of the other percentiles are more accurate as well. The standard deviation estimates are closer to 2.161, as expected. We cannot reject the null hypothesis that these data sets come from a lognormal distribution, based on the KSL test.

Distributions

1ln1000



LogNormal(0,1)

Quantiles

100.0%	maximum	22.171
99.5%		13.843
97.5%		7.187
90.0%		3.600
75.0%	quartile	1.965
50.0%	median	0.999
25.0%	quartile	0.509
10.0%		0.276
2.5%		0.139
0.5%		0.074
0.0%	minimum	0.029

Moments

Mean	1.640772
Std Dev	2.0538219
Std Err Mean	0.0649476
upper 95% Mean	1.7682212
lower 95% Mean	1.5133227
N	1000

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	1.640772	1.513323	1.768221	0.950
Std Dev	2.053822	1.967587	2.148021	

Fitted LogNormal

Fixed Parameters

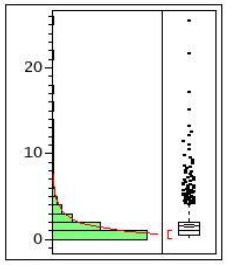
Parameter	Value	Fixed	Lower 95%	Upper 95%
Mu	0.000000	*	.	.
Sigma	1.000000	*	.	.

Goodness-of-Fit Test

KSL Test

D		Prob>D
0.001369	>	0.1500

2ln1000



LogNormal(0,1)

Quantiles

100.0%	maximum	25.589
99.5%		13.220
97.5%		7.207
90.0%		3.612
75.0%	quartile	1.962
50.0%	median	0.999
25.0%	quartile	0.509
10.0%		0.277
2.5%		0.140
0.5%		0.073
0.0%	minimum	0.039

Moments

Mean	1.6456627
Std Dev	2.1022669
Std Err Mean	0.0664795
upper 95% Mean	1.7761183
lower 95% Mean	1.5152072
N	1000

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	1.645663	1.515207	1.776118	0.950
Std Dev	2.102267	2.013998	2.198688	

Fitted LogNormal

Fixed Parameters

Parameter	Value	Fixed	Lower 95%	Upper 95%
Mu	0.000000	*	.	.
Sigma	1.000000	*	.	.

Goodness-of-Fit Test

KSL Test

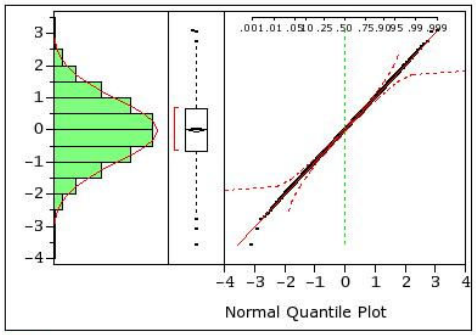
D		Prob>D
0.001366	>	0.1500

Figure 13. Summary Statistics, Lognormal Distribution, N = 1000

The log-transform of the 1000 point data sets follows a normal distribution very closely, as shown below in Figure 14:

Distributions

log(1ln1000)



Normal(0,1)

Quantiles

100.0%	maximum	3.099
99.5%		2.628
97.5%		1.972
90.0%		1.281
75.0%	quartile	0.675
50.0%	median	-0.001
25.0%	quartile	-0.676
10.0%		-1.286
2.5%		-1.971
0.5%		-2.598
0.0%	minimum	-3.550

Moments

Mean	-0.001363
Std Dev	1.0008375
Std Err Mean	0.0316493
upper 95% Mean	0.060744
lower 95% Mean	-0.063469
N	1000

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	-0.00136	-0.06347	0.060744	0.950
Std Dev	1.000837	0.958815	1.046741	

Fitted Normal

Fixed Parameters

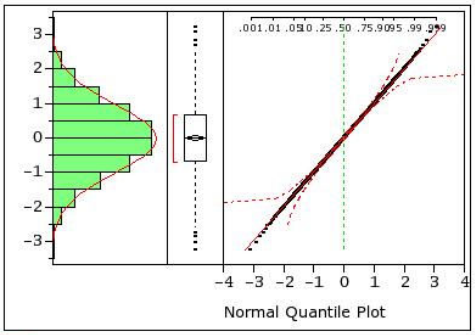
Parameter	Value	Fixed	Lower 95%	Upper 95%
Mu	0.000000	*	.	.
Sigma	1.000000	*	.	.

Goodness-of-Fit Test

KSL Test

D		Prob>D
0.001369	>	0.2500

log(2n1000)



Normal(0,1)

Quantiles

100.0%	maximum	3.242
99.5%		2.582
97.5%		1.975
90.0%		1.284
75.0%	quartile	0.674
50.0%	median	-0.0008
25.0%	quartile	-0.676
10.0%		-1.284
2.5%		-1.964
0.5%		-2.621
0.0%	minimum	-3.242

Moments

Mean	-0.000934
Std Dev	1.0006722
Std Err Mean	0.031644
upper 95% Mean	0.0611621
lower 95% Mean	-0.063031
N	1000

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	-0.00093	-0.06303	0.061162	0.950
Std Dev	1.000672	0.958656	1.046568	

Fitted Normal

Fixed Parameters

Parameter	Value	Fixed	Lower 95%	Upper 95%
Mu	0.000000	*	.	.
Sigma	1.000000	*	.	.

Goodness-of-Fit Test

KSL Test

D		Prob>D
0.001366	>	0.2500

Figure 14. Summary Statistics, Log-transformed Lognormal Distribution, N = 1000

Summary Statistics and Graphical Comparison: N = 10000

We perform a similar transformation for N=10000: first we look at the raw data, then log-transform it. The raw data has a lognormal distribution shape, as shown in Figure 15. Note that in these samples, the maximum value is in the mid-forties. Again, this is an example of getting better samples in the tails for long-tailed distributions as you increase the number of samples. The 99.95th percentile of the true distribution is 26.86. For 10000 samples, we expect 5 samples to lie above 26.86. Sample set 1ln10000 and sample set 2ln20000 both have exactly 5 samples above this value. The 97.5th percentile estimate is more accurate than that generate in the 1000-point sample sets, as expected. Most of the other percentiles are more accurate as well. The standard deviation estimates are closer to 2.161, as expected. We cannot reject the null hypothesis that these data sets come from a lognormal distribution, based on the KSL test. Note that as the number of samples increases from 100 to 10000, we see the means and standard deviations from the sample sets converging to the true values as shown in Table 2:

Sample Size	Case 1	Case 2	True Mean
100	1.616	1.713	1.648
1000	1.641	1.646	1.648
10000	1.647	1.647	1.648
			True Std. Dev.
100	1.875	2.474	2.161
1000	2.054	2.102	2.161
10000	2.135	2.133	2.161

Table 2. Summary, Lognormal Distribution

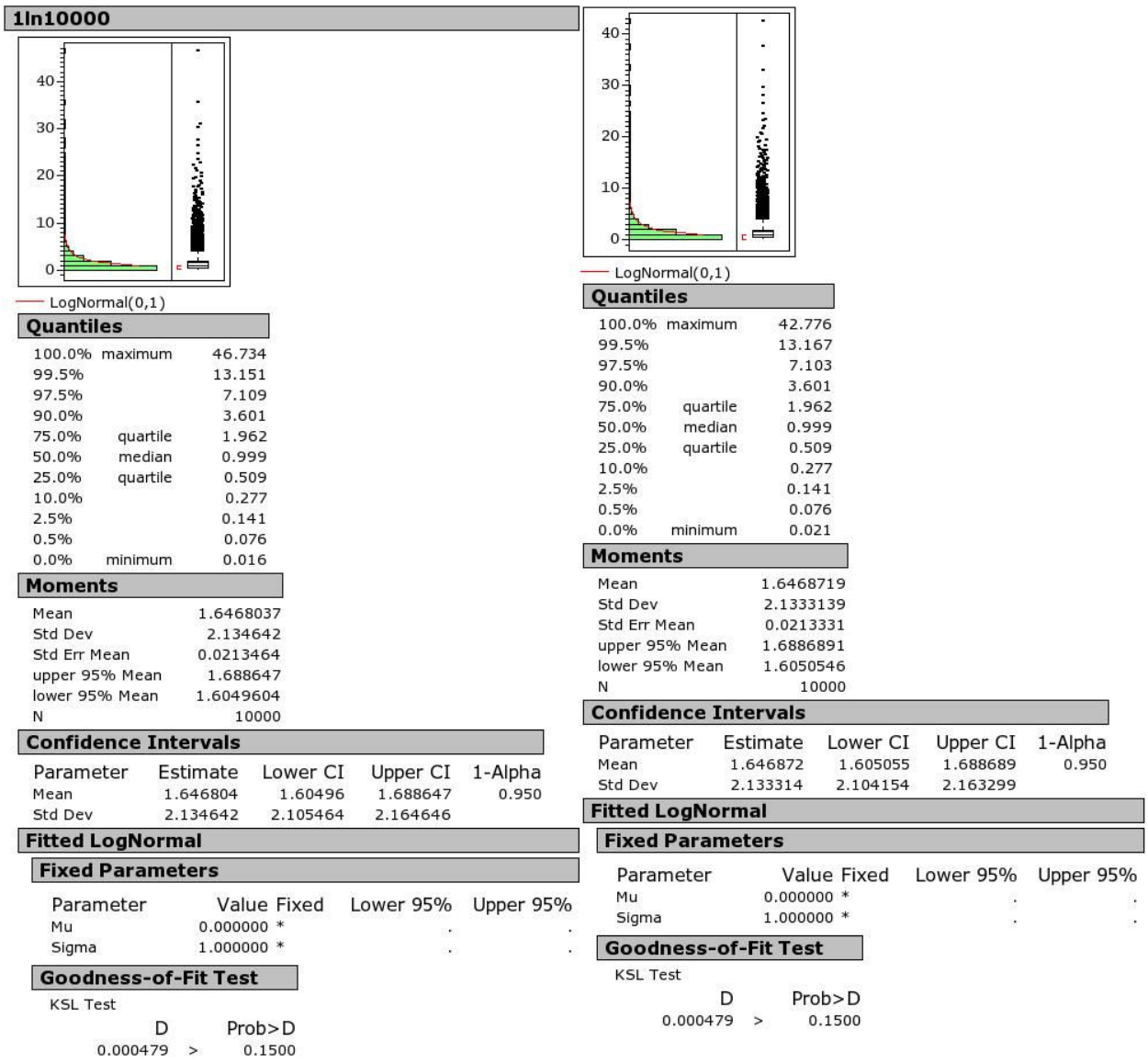


Figure 15. Summary Statistics, Lognormal Distribution, N = 10000

Figure 16 shows the log-transform for the 10000-point data sets. Note that these data sets strongly support the hypothesis that the underlying distribution is normal.

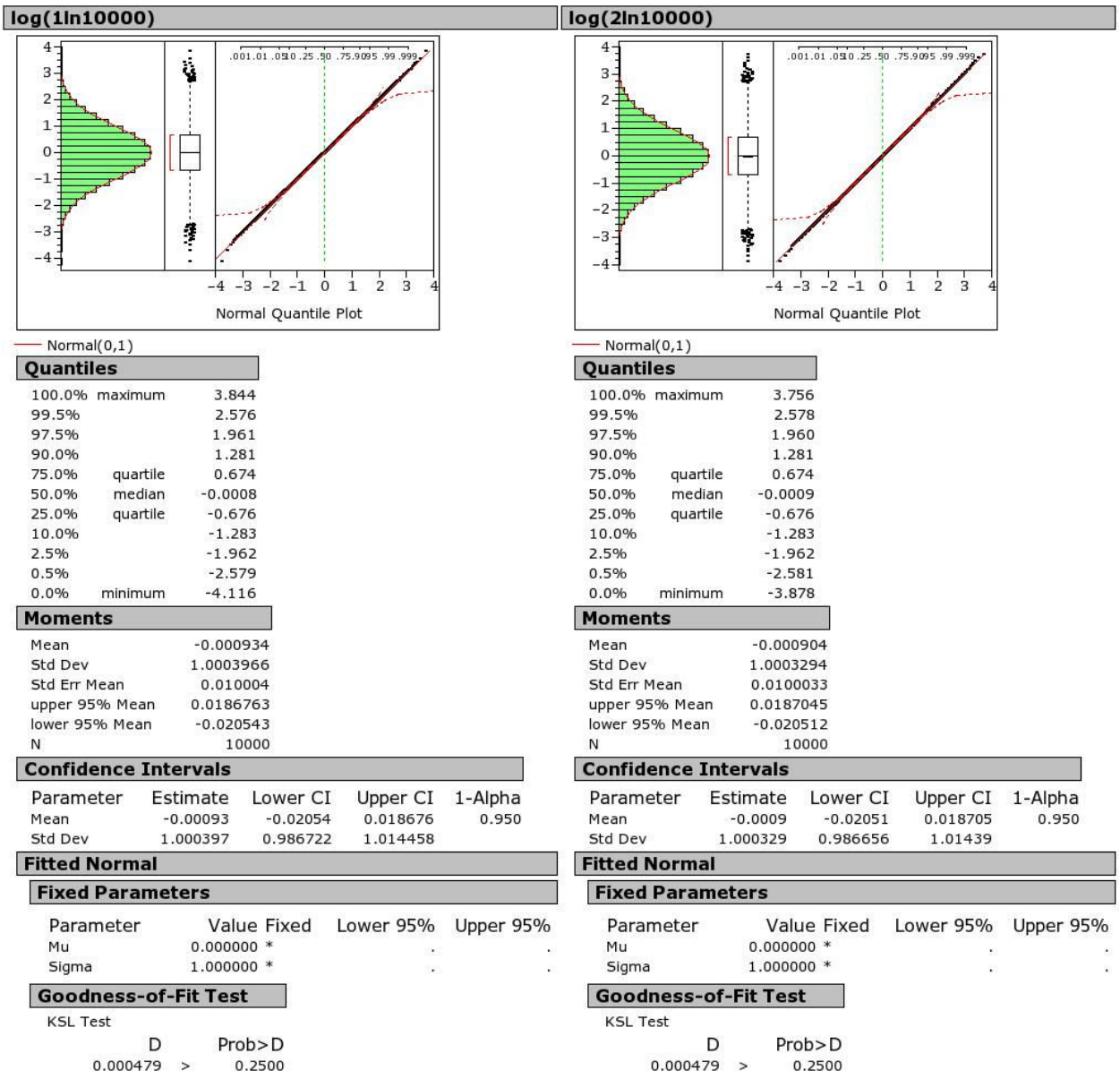


Figure 16. Summary Statistics, Log-transformed Lognormal Distribution, N = 10000

6. The Uniform Distribution

The LHS software implemented in DAKOTA provides the user with one method for sampling from the uniform distribution. The uniform distribution is defined by the density function

$$f(x) = \frac{1}{U - L} \quad U > L$$

where U and L denote the uppder and lower bounds of the uniform distribution, respectively. The mean of the uniform distribution is given by: $\frac{U+L}{2}$, and the standard deviation is given by: $\sqrt{\frac{(U+L)^2}{12}}$.

For the purposes of the V&V analysis of the uniform distribution in LHS, three runs of the LHS code were performed in DAKOTA. Each run involved 2 uniformly distributed, uncorrelated random variables. Each random variable was chosen from the uniform distribution with a lower bound of zero and an upper bound of one. The mean of this distribution is 0.5, and the standard deviation is 0.2887. As before, the first run had 100 samples, the second run had 1000 samples, and the third run had 10,000 samples. The sections below provide results of testing with these sample data sets.

Summary Statistics and Graphical Comparison: N = 100

Figure 17 shows that the samples generated for a uniform with bounds [0,1] do follow a uniform distribution. The mean is extremely close to 0.50 for both 1u100 and 2u100. The standard deviation is also close to 0.2887 in both cases. The histogram shows that there are the same number of points (10) in each bin [0,0.1) [0.1,0.2) etc. The stem-and-leaf graph is a little misleading because of the way JMP does its rounding. For example, in case 2, the largest value on the stem and leaf graph looks like there is a sample at 1.00. There is not a sample value of 1.00, but instead, the largest value of this sample, 0.9987, is rounded up to 1.0 for the stem and leaf plot.

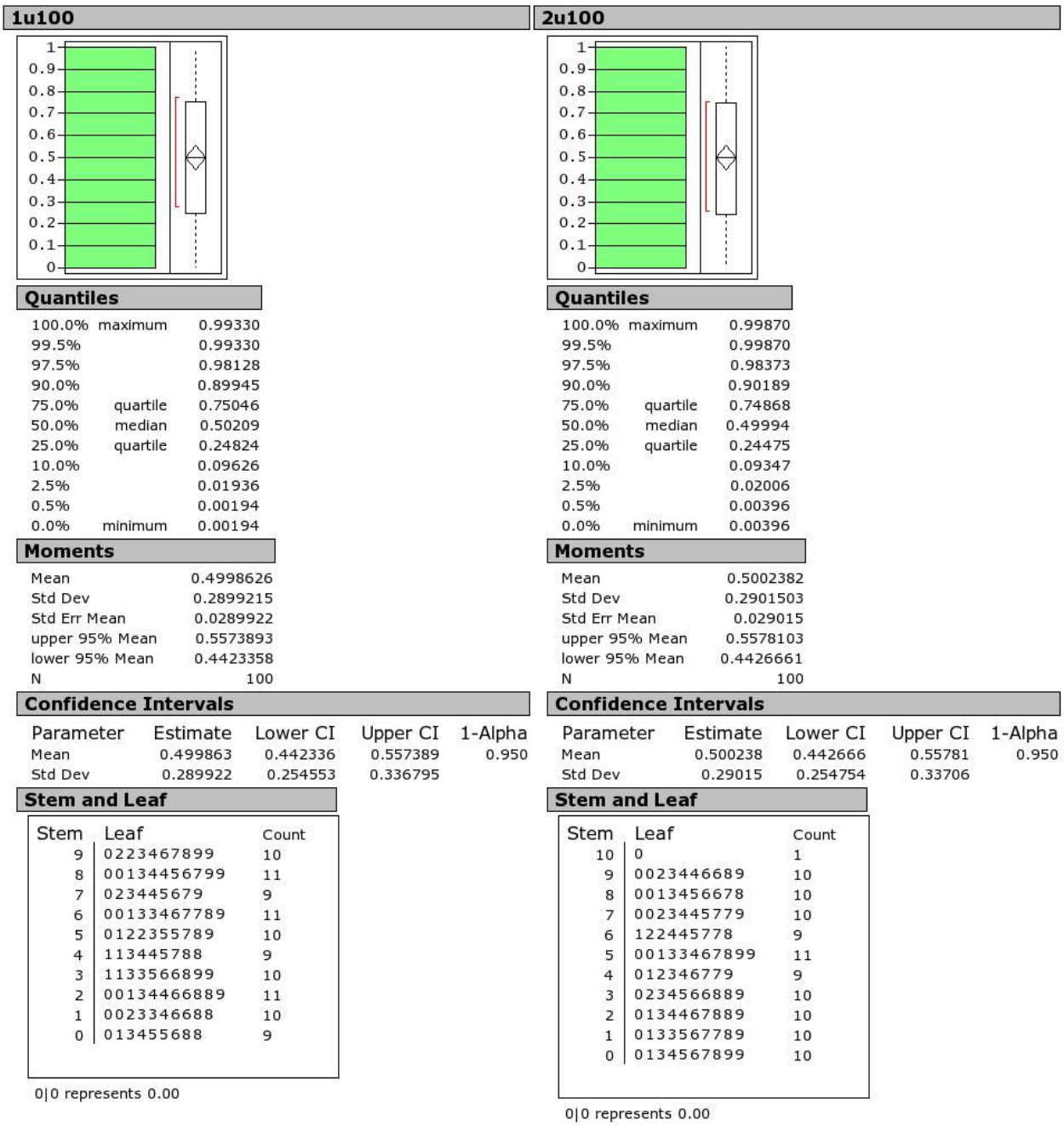


Figure 17. Summary Statistics, Uniform Distribution, N = 100

Formal statistical tests: N = 100

Note that it is possible to apply some of the formal goodness-of-fit hypothesis testing to uniform distributions. However, the vast majority of the tests supported in commercial software are specifically designed for testing if a distribution is normal. Neither Minitab nor JMP directly

supports testing for uniformity. However, we developed the test statistics based on the sample values and performed the analysis.

Recall that the Chi-square test measures the difference between the expected proportion of samples that will fall in the j^{th} interval, p_j, and the actual number of samples that falls in the jth interval, N_j:

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - Np_j)^2}{Np_j}$$

where N is the total number of samples.

Because LHS is stratified, this test statistic will be zero. For example, if we divide the 100 samples up into 10 bins, we expect that 10 samples will fall within the first bin [0, 0.1), 10 samples will fall into the second bin [0.1, 0.2), etc. The results from these two LHS samples show that there are indeed exactly 10 samples in the first bin, 10 samples in the second bin, etc.

The results of the χ^2 test strongly support that the null hypothesis of a uniform distribution cannot be rejected. Below is the Minitab output for this Chi-square test. Note that the p-value is 1.0, meaning that the probability that one would obtain these results with a uniform distribution is essentially 1.0.

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: 1u100

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
1	10	0.1	10	0
2	10	0.1	10	0
3	10	0.1	10	0
4	10	0.1	10	0
5	10	0.1	10	0
6	10	0.1	10	0
7	10	0.1	10	0
8	10	0.1	10	0
9	10	0.1	10	0
10	10	0.1	10	0

N	DF	Chi-Sq	P-Value
100	9	0	1.000

Because χ^2 test statistic will always be zero for uniform samples generated with LHS, pointing to accepting the null hypothesis, I implemented one other test statistic as an additional verification check.

Recall that the Kolmogorov-Smirnov test statistic, D, is given by:

$$D = \max_{(1 \le i \le N)} \left(F(X_{(i)}) - \frac{i-1}{N}, \frac{i}{N} - F(X_{(i)}) \right)$$

Where one will reject the null hypothesis if: $\left(\sqrt{N}+0.12+\frac{0.11}{\sqrt{N}}\right)D > c_{1-\alpha}$, where the value of $c_{1-\alpha}$ is 1.38 when $\alpha = 0.05$. For data set 1u100, D = 0.00999 and $\left(\sqrt{N}+0.12+\frac{0.11}{\sqrt{N}}\right)D = .1013$, so we cannot reject the null hypothesis of a uniform distribution. For data set 2u100, D = 0.00997, and $\left(\sqrt{N}+0.12+\frac{0.11}{\sqrt{N}}\right)D = .1010$ and again we cannot reject the null hypothesis.

Summary Statistics and Graphical Comparison: N = 1000

Figure 18 shows that the samples generated for a uniform with bounds [0,1] do follow a uniform distribution. The mean is extremely close to 0.50 for both 1u1000 and 2u1000. The standard deviation is also close to 0.2887 in both cases. The histogram shows that there are the same number of points (100) in each bin [0,0.1) [0.1,0.2) etc. The stem-and-leaf plot was not printed because it was too large with this many points at the resolution JMP output.

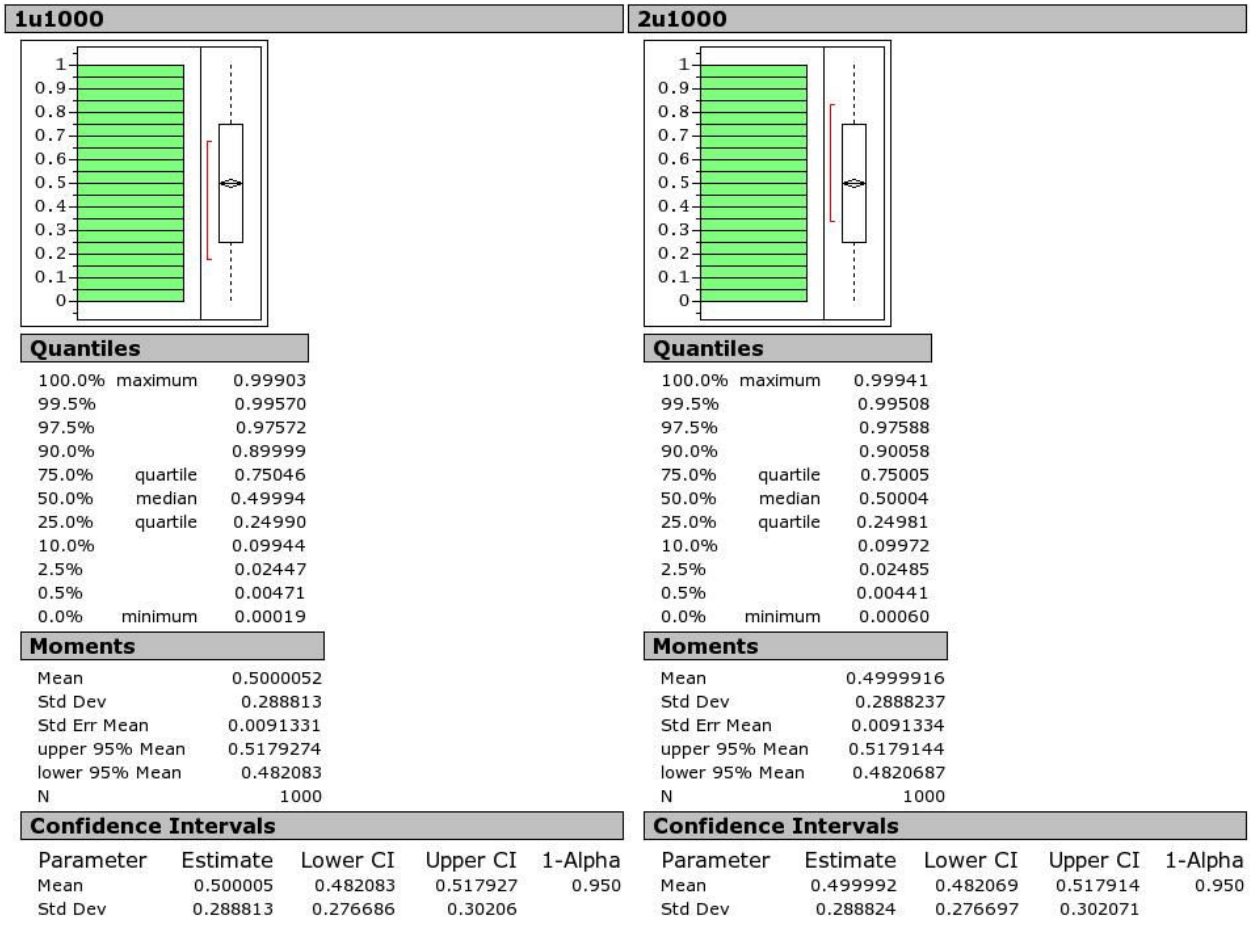


Figure 18. Summary Statistics, Uniform Distribution, N = 1000

Formal statistical testing is not very useful in the case of the Chi-square test, since the Chi-square test statistics is zero and again we accept the null hypothesis that the data come from a uniform distribution. For the K-S test, the test statistic D is 0.001 for both samples. For both data sets 1u1000 and 2u1000 , D = 0.001and $\left(\sqrt{N}+0.12+\frac{0.11}{\sqrt{N}}\right)D=.0317$, so we cannot reject the null hypothesis of a uniform distribution.

Summary Statistics and Graphical Comparison: N = 10000

Figure 19 shows that the samples generated for a uniform with bounds [0,1] do follow a uniform distribution. The mean is extremely close to 0.50 for both 1u10000 and 2u10000. The standard deviation is also close to 0.2887 in both cases. The histogram shows that there are the same number of points (100) in each bin [0,0.1) [0.1,0.2) etc.

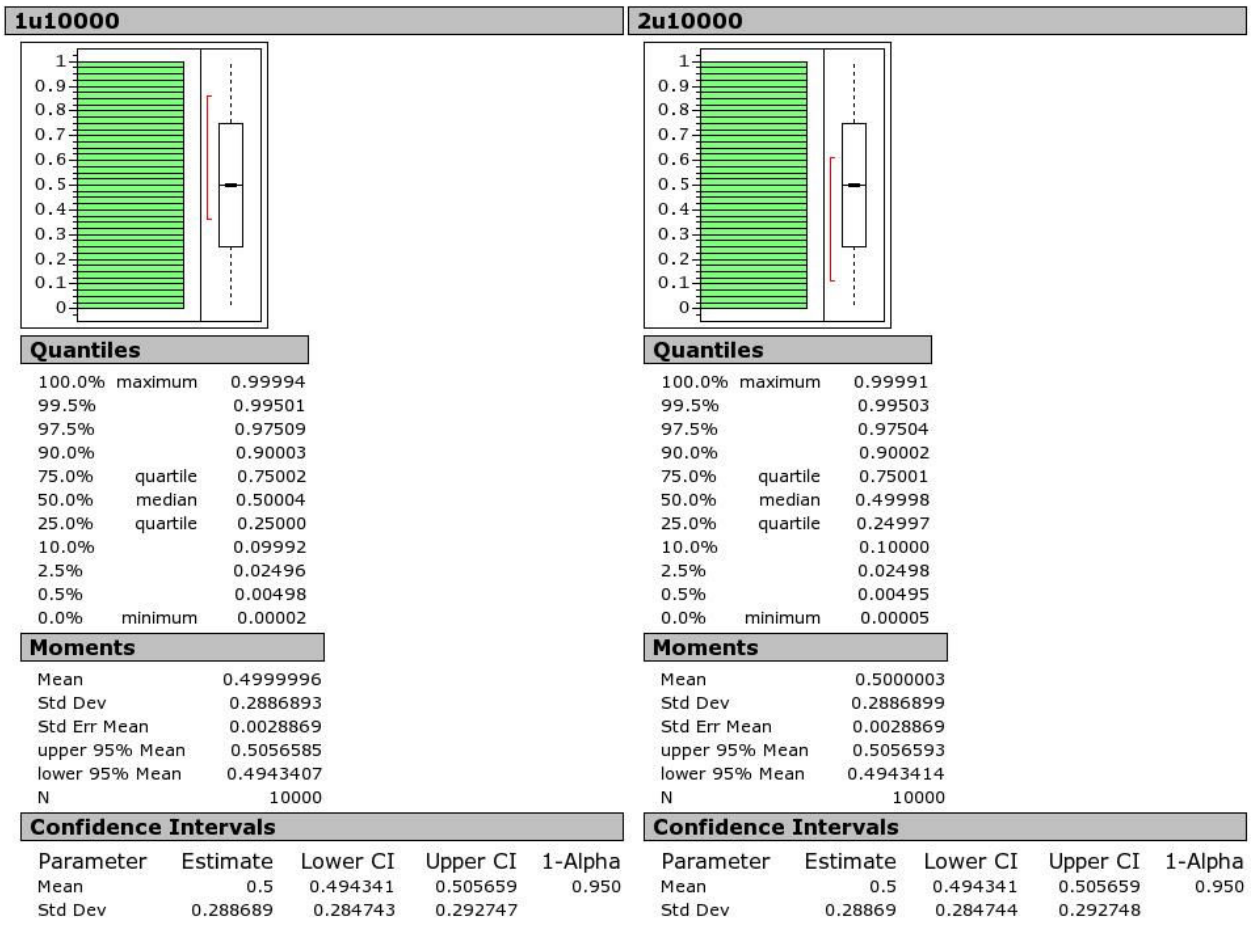


Figure 19. Summary Statistics, Uniform Distribution, N = 100

Formal statistical testing is again not very useful in the case of the Chi-square test, since the Chi-square test statistics is zero and again we accept the null hypothesis that the data come from a uniform distribution. For the K-S test, the test statistic D is 0.0001 for both samples. For both data sets 1u1000 and 2u1000 , D = 0.001and $\left(\sqrt{N}+0.12+\frac{0.11}{\sqrt{N}}\right)D = .010$, so we cannot reject the null hypothesis of a uniform distribution.

7. Large Scale Tests

The test results presented above for normal, lognormal, and uniform distributions were small tests: only 2 variables of each distribution type were sampled. To mimic the needs of the ASC milestones and also to ensure more robustness in our verification tests, we ran some larger verification tests. The test results are shown in Table 3.

The test set-up is as follows: we produced joint samples of 30 or 50 variables simultaneously, with sample sizes ranging from 50 to 10000 samples. This means that we produced a joint sample for 30 normal random variables, for example, and not 30 normal random variables each sampled individually. Often, in Monte Carlo sampling, there are fairly large correlations between input variable sample values (e.g., variable 16 may be correlated with variable 23 with a correlation coefficient of .4 or even higher). If these random variables are independent which is often the case, one would NOT like to have high correlation values in the sample data. We used the restricted pairing method developed by Iman and Conover to specify that zero or near-zero correlation be induced between the sample variable values. The restricted pairing method worked extremely well. For example, in the sample set of 10000 samples, the correlations between various pairs of input variables were on the order of 10^{-4}.

The testing showed that all of the verification tests for normal, lognormal, and uniform distributions passed the Kolmogorov-Smirnov test. That is, we cannot reject the hypothesis that these 30 or 50 random variables generated each come from a normal, lognormal, or uniform distribution, respectively. The test results also show that the summary statistics converge as the sample size increases, which is what we expect. As sample size goes from 50 to 10,000, we see that the average sample mean and average sample standard deviation approach the true mean and standard deviation for each of the three distribution types.

	Number of Random Vars	Sample Size	K-S Test	Average mean	True Mean	Average Std.Deviation	True Std. Deviation
Normal	30	50	All pass	0.000149	0	1.016044	1
	50	100	All pass	-0.000310	0	1.008588	1
	50	1000	All pass	-0.000015	0	1.000501	1
	50	10000	All pass	-0.000001	0	1.000047	1
Lognormal	30	50	All pass	1.664300	1.647821	2.147719	2.161197
	50	100	All pass	1.651513	1.647821	2.117042	2.161197
	50	1000	All pass	1.649248	1.647821	2.173071	2.161197
	50	10000	All pass	1.647750	1.647821	2.157194	2.161197
Uniform	30	50	All pass	0.499936	0.500000	0.291800	0.288675
	50	100	All pass	0.500027	0.500000	0.290231	0.288675
	50	1000	All pass	0.500000	0.500000	0.288823	0.288675
	50	10000	All pass	0.500000	0.500000	0.288689	0.288675

Table 3. Test Results for Large Scale Verification Tests

8. Summary

This document provides verification test results for normal, lognormal, and uniform distributions that are used in Sandia's Latin Hypercube Sampling (LHS) software as accessed through DAKOTA. The purpose of this testing is to verify that the sample values being generated in LHS are distributed according to the desired distribution types. The testing of distribution correctness is done by examining summary statistics, graphical comparisons using quantile-quantile plots, and format statistical tests such as the Chi-square test, the Kolmogorov-Smirnov test, and the Anderson-Darling test. The overall results from the testing indicate that the generation of normal, lognormal, and uniform distributions in LHS as accessed through DAKOTA is acceptable. LHS has been a powerful tool for sampling statistical distributions in uncertainty analyses for more than 20 years. The LHS UNIX version that is implemented in DAKOTA represents an investment to modernize the code capabilities and allow this valuable uncertainty analysis capability to remain viable for large-scale simulation models running under a Linux or UNIX operating system.

9. References

Eldred, M.S., Giunta, A. A., van Bloemen Waanders, B.G., Wojtkiewicz, S.F., Hart, W.E., and Alleva, M.P. (2001). "DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis." Technical Reports SAND2001-3796 (User's Manual), SAND2001-3515 (Reference Manual), and SAND2001-3514 (Developer's Manual). Sandia National Laboratories, Albuquerque NM.

Helton, J. C. and Davis, F. J. (2001). "Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems." Technical Report SAND2001-0417, Sandia National Laboratories, Albuquerque, NM.

Iman, R.L., and Conover, W.J. (1982a). "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables," *Communications in Statistics*, B11(3), 311-334.

Iman, R.L., Davenport, J.M., and Ziegler, D.K. (1980). "Latin Hypercube Sampling (Program User's Guide)," Technical Report SAND79-1473, Sandia National Laboratories, Albuquerque, NM.

Law, A.M. and W.D. Kelton (1997). *Simulation Modeling and Analysis*, 2nd edition. McGraw-Hill.

McKay, M.D., Conover, W.J., and Beckman, R.J. (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," Technometrics, 221, 239-245.

NIST/SEMATECH *e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, January 2006.

Swiler, L. P. and G. D. Wyss. "A User's Guide to Sandia's Latin Hypercube Sampling Software: LHS Unix Library/Standalone Version." Technical Report SAND2004-2439. Sandia National Laboratories, Albuquerque NM.

Wyss, G. D., and Jorgensen, K. H. "A User's Guide to LHS: Sandia's Latin Hypercube Sampling Software." Technical Report SAND1998-0210. Sandia National Laboratories, Albuquerque, NM.

DISTRIBUTION

MS0757	G. D. Wyss, 6442
MS0828	A. A. Giunta, 1533
MS0779	J.C. Helton, 1533
MS0828	M. Pilch, 1533
MS0828	V. J. Romero, 1533
MS0828	W. L. Oberkampf, 1533
MS0370	S. L. Brown, 1411
MS0370	M. S. Eldred, 1411
MS0370	S. A. Mitchell, 1411
MS0370	L. P. Swiler, 1411 [2]
MS1011	J. E. Campbell, 6642
MS9018	Central Technical Files, 8945-1 [2]
MS0899	Technical Library, 4536 [2]